Second Language Pronunciation Assessment

SECOND LANGUAGE ACQUISITION

***Series Editors*:** Professor David Singleton, *University of Pannonia, Hungary* and Fellow Emeritus, *Trinity College, Dublin, Ireland* and Dr Simone E. Pfenninger, *University of Salzburg, Austria*

This series brings together titles dealing with a variety of aspects of language acquisition and processing in situations where a language or languages other than the native language is involved. Second language is thus interpreted in its broadest possible sense. The volumes included in the series all offer in their different ways, on the one hand, exposition and discussion of empirical findings and, on the other, some degree of theoretical reflection. In this latter connection, no particular theoretical stance is privileged in the series; nor is any relevant perspective – sociolinguistic, psycholinguistic, neurolinguistic, etc. – deemed out of place. The intended readership of the series includes final-year undergraduates working on second language acquisition projects, postgraduate students involved in second language acquisition research, and researchers, teachers and policy-makers in general whose interests include a second language acquisition component.

Full details of all the books in this series and of all our other publications can be found on http://www.multilingual-matters.com, or by writing to Multilingual Matters, St Nicholas House, 31–34 High Street, Bristol BS1 2AW, UK.

SECOND LANGUAGE ACQUISITION: 107

Second Language Pronunciation Assessment

Interdisciplinary Perspectives

Edited by
Talia Isaacs and Pavel Trofimovich

MULTILINGUAL MATTERS
Bristol • Blue Ridge Summit

In Memory of Alan Davies and Danielle Guénette

DOI 10.21832/ISAACS6848
Library of Congress Cataloging in Publication Data
A catalog record for this book is available from the Library of Congress.
Names: Isaacs, Talia, editor. | Trofimovich, Pavel, editor.
Title: Second Language Pronunciation Assessment: Interdisciplinary Perspectives/Edited by Talia Isaacs and Pavel Trofimovich.
Description: Bristol: Multilingual Matters, [2017] | Series: Second Language Acquisition: 107 | Includes bibliographical references and index.
Identifiers: LCCN 2016031375 | ISBN 9781783096848 (hbk : alk. paper) | ISBN 9781783096831 (pbk : alk. paper) | ISBN 9781783096879 (kindle)
Subjects: LCSH: Second language acquisition—Ability testing. | Language and languages—Pronunciation—Ability testing. | Language and languages—Pronunciation for foreign speakers. | Language and languages—Study and teaching—Foreign speakers. | Second language acquisition—Research.
Classification: LCC P118.75 .S43 2015 | DDC 418.0076—dc23 LC record available at https://lccn.loc.gov/2016031375

British Library Cataloguing in Publication Data
A catalogue entry for this book is available from the British Library.

ISBN-13: 978-1-78309-684-8 (hbk)
ISBN-13: 978-1-78309-683-1 (pbk)
ISBN-13: 978-1-78309-685-5 (pdf)
ISBN-13: 978-1-78309-686-2 (epub)

Multilingual Matters
UK: St Nicholas House, 31–34 High Street, Bristol BS1 2AW, UK.
USA: NBN, Blue Ridge Summit, PA, USA.

Website: www.multilingual-matters.com
Twitter: Multi_Ling_Mat
Facebook: https://www.facebook.com/multilingualmatters
Blog: www.channelviewpublications.wordpress.com

Typeset by Nova Techset Private Limited, Bengaluru and Chennai, India.
Printed and bound in the UK by the CPI Books Group Ltd.
Printed and bound in the US by Edwards Brothers Malloy, Inc.

Contents

Acknowledgements

This edited volume, which brings together different but complementary research perspectives to establish a common platform in which to discuss issues relevant to assessing second language (L2) pronunciation, would not have been possible without the contributions and commitment of the authors, who explore key issues through different disciplinary lenses in the chapters that make up this volume. The vision for the book arose during a cold Canadian winter at the beginning of the second decade of the 21st century, when a sense of momentum for interdisciplinary research on L2 pronunciation assessment was palpable and, indeed, has been growing in the years since. It is a joy to bring together emergent thinking in a single volume in what we hope will be an indispensable point of reference for researchers and practitioners wishing to read up on and undertake further work in this area.

There were some unforeseen challenges in the process of pulling this volume together. During the period between the authors' initial chapter submission deadline in early 2015 and the submission of the entire manuscript to the publisher by the end of the calendar year, sadly, two book contributors passed away. Alan Davies was a monumental and inspirational figure in the field of language assessment for generations of researchers. News of his loss on language testing and applied linguistics mailing lists was accompanied by an outpouring of tributes from former students and colleagues around the globe. Of the many applied linguistics topics with a social bent that Alan wrote about prolifically, his scholarship on the native speaker is among the most noteworthy. Alan's chapter included in this edited volume, written just over six months before his passing, is, in some places, reminiscent of an armchair conversation. His voice is clear and his ideas will continue to resonate for generations to come.

We were also touched by the untimely death of Danielle Guénette, co-author of the chapter on the topic of teachers' assessments of French lingua franca interactions with Sara Kennedy and Josée Blanchet. As the lead author attested, Danielle was instrumental to data collection and data processing in that study. Danielle had an infectious positivity and *joie de vivre* and her passion for language teaching permeated many of her interactions. It is an

honour to be able to dedicate this book to the memory of our two most worthy colleagues.

We are extremely grateful to our many students, collaborators and intellectual sounding boards, whose passion and thirst for research through over a decade of conversations has inspired the content of this volume. These individuals are numerous and continue to shape our thinking. In relation to the production of this volume specifically, we would like to acknowledge Sohaib Sandhu for his assistance in preparing an Appendix to our book proposal and particularly Kym Taylor Reid for her help with copyediting the entire volume. Any remaining errors are our own.

We also sincerely thank Laura Longworth, Tommi Grover and the whole team at Multilingual Matters/Channel View Publications for their enthusiasm about the topic, congeniality, professionalism, prompt responses to our queries, and openness to the prospect of pursuing open access, allowing this volume to break new ground and reach a wider audience as intended. It is a rare treat to have both a local (Bristol-based) and world-class publisher with a track record of working with high-calibre researchers at our doorstep, and we are so pleased to have capitalized on this opportunity. We are also grateful to David Singleton and Simone Pfenninger, our series editors, for their rapid review of the manuscript and insightful comments. Finally, we acknowledge grants from both the FP7 Marie Skłodowska-Curie Actions (PCIG10-GA-2011-303413), and the Social Sciences and Humanities Research Council of Canada, which supported the preparation of this edited collection, and for funding from the European Commission OpenAIRE FP7 Post-Grant Open Access Pilot, enabling us to make this manuscript publically available.

Most of all, we thank Padraig and Sarita and Katya, whose immensely positive effect on our lives is difficult to express through language (even for applied linguists) but very deeply felt.

Talia Isaacs and Pavel Trofimovich
December 2015

Contributors

Laura Ballard is a doctoral student in the Second Language Studies Program at Michigan State University, USA. She is a contributor to various ESL assessment projects in the Testing Office at Michigan State University's English Language Center. She researches ESL assessment and language testing policy issues.

Fiona Barker has a teaching background and a PhD in Corpus Linguistics (Cardiff, UK). She trains and publishes internationally on aspects of English learning, teaching and assessment, focusing on action research and assessment literacy for practitioners and the uses of technology for English language learning and assessment.

Josée Blanchet is a tenured Lecturer at the Language School in the Faculty of Communication at the Université du Québec à Montréal (UQAM), Canada. Her research focuses on listening and pronunciation instruction in L2 French. She also investigates intercultural practices in second language instruction.

Kevin Browne is an Associate Professor of English at Yamanashi Prefectural University in Japan, and is currently completing a doctorate in Language Testing at the University of Leicester. He received an MA in Applied Linguistics from the University of Melbourne and a BA in English from Loyola University New Orleans.

Isabelle Darcy is an Associate Professor in the Department of Second Language Studies at Indiana University, USA. She obtained a PhD in Linguistics and Cognitive Science from the EHESS in Paris (France) and from the Gutenberg University in Mainz (Germany). Her research includes native and nonnative phonological acquisition, speech perception and word recognition.

Alan Davies, who passed away in September 2015 prior to the completion of this edited volume, was Professor Emeritus of Applied Linguistics at the University of Edinburgh, UK, where he was initially appointed in 1965.

Among his many areas of expertise, Alan is particularly well known for his extensive outputs problematizing the concept of the native speaker, perhaps the last of which appears in this volume.

Glenn Fulcher is a Professor of Education and Language Assessment in the School of Education, University of Leicester, UK. Recent books include the Routledge *Handbook of Language Testing*, *Practical Language Testing* and *Language Testing Re-examined: A Philosophical and Social Inquiry*. His website (http://languagetesting.info) is widely used in teaching and researching language assessment.

Evelina Galaczi is Principal Research Manager at Cambridge English, University of Cambridge, UK. Her research focuses primarily on speaking assessment and the role of assessment to support learning (Learning Oriented Assessment). She regularly presents at international conferences and has published in academic forums including *Applied Linguistics, Language Assessment Quarterly*, *Assessment in Education*, and the Studies in Language Assessment series (CUP).

Danielle Guénette was an Associate Professor in the Département de Didactique des Langues in the Faculté des Sciences de l'Education at the Université du Québec à Montréal (UQAM), Canada. She passed away in February 2015 after an illness. She taught and conducted research on L2 speech, written corrective feedback and L2 teacher education.

Luke Harding is a Senior Lecturer in the Department of Linguistics and English Language at Lancaster University, UK. His research is mainly in the area of language testing, specifically listening assessment, pronunciation and intelligibility, and the challenges of World Englishes and English as a lingua franca for language assessment.

Talia Isaacs is a Senior Lecturer in Applied Linguistics and TESOL at the UCL Centre for Applied Linguistics, UCL Institute of Education, University College London, UK. Her research investigates learners' performances and raters' judgments of L2 speech (particularly pronunciation). She serves on the editorial boards of *Language Assessment Quarterly*, *Language Testing* and *The Journal of Second Language Pronunciation*.

Sara Kennedy is an Associate Professor in the Department of Education at Concordia University in Montreal, Canada. She teaches and conducts research on the teaching, learning, assessment and use of second language speech, with a particular interest in L2 pronunciation.

Ute Knoch is the Director of the Language Testing Research Centre at the University of Melbourne, Australia. Her research interests are in the area of

writing assessment and assessing languages for academic and professional purposes. In 2014 she was awarded the TOEFL Outstanding Young Scholar Award by the Educational Testing Service.

Aike Li obtained her PhD from the University of Cambridge, studying second language development of prosody. She is now a Lecturer at the Communication University of China, teaching English Phonetics. Her research areas include second language acquisition, phonetics and phonology, as well as speech communication.

Stephanie Lindemann is an Associate Professor of Applied Linguistics at Georgia State University, USA. Her research focuses on the native speaker role in communication with nonnative speakers, including perceptions of nonnative speech and attitudes towards such speech. She is currently investigating ways of improving attitudes and comprehension of nonnative speech.

Joan C. Mora is an Associate Professor in the English Department at the University of Barcelona, Spain. His research examines the role of input and aptitude in L2 phonological acquisition and the effects of learning context and individual differences in the development of L2 pronunciation and oral fluency in instructed SLA.

Brechtje Post is a Reader in Experimental Phonology in the Department of Theoretical and Applied Linguistics of the University of Cambridge, UK. Her research focuses primarily on prosody, which she investigates from a phonetic, phonological, acquisitional, cognitive and neural perspective. She publishes in journals such as *Cognition*, *Frontiers in Psychology*, *Language and Speech*, *Langue Française*, *Journal of Phonetics* and *Studies in Second Language Acquisition*.

Kazuya Saito is a Lecturer in Second Language Learning at the Department of Applied Linguistics and Communication, Birkbeck University of London, UK. His research investigates how instruction and corrective feedback can help adult learners develop their L2 oral proficiency, especially in the domains of pronunciation, listening, vocabulary and grammar.

Elaine Schmidt obtained her PhD from the University of Cambridge, working on the prosodic development of bilingual children. She is now a Postdoctoral Research Fellow in the Child Language Lab at Macquarie University (Sydney, Australia) and an Associate Investigator at the ARC Centre of Excellence in Cognition and its Disorders. Her research focuses on prosodic processing in children and adults through behavioural, EEG and MEG experiments.

Andrew Sewell is an Assistant Professor in the Department of English at Lingnan University, Hong Kong. He has extensive experience of language teaching in Asia, and has worked as an examiner for several international examinations. His interdisciplinary research interests include linguistic, sociolinguistic and pedagogical aspects of English as an international language.

Paul D. Toth is an Associate Professor of Spanish Applied Linguistics at Temple University, USA. His research on task-based instruction has twice received the ACTFL/MLJ Pimsleur Award, and has appeared in the 2011 *Best of Language Learning* volume. He is currently interested in how metalinguistic knowledge and discourse pragmatics affect L2 development.

Pavel Trofimovich is a Professor of Applied Linguistics in the Department of Education at Concordia University, Canada. His research focuses on cognitive aspects of second language (L2) processing, phonology, sociolinguistic aspects of L2 acquisition, and the teaching of L2 pronunciation. He is the current editor of *Language Learning*.

Elvis Wagner is an Associate Professor of TESOL at Temple University, USA. His current research focuses on how L2 listeners process and comprehend unscripted, spontaneous spoken language, and how this type of language differs from the scripted spoken texts learners are often exposed to in the L2 classroom.

Stuart Webb is a Professor in the Faculty of Education at Western University, Canada. His research interests include teaching and learning vocabulary, second language acquisition, and extensive reading and listening.

Paula Winke is an Associate Professor in the Department of Linguistics, Germanic, Slavic and Asian Languages at Michigan State University, USA. She researches language testing and language teaching methods, as well as attention in task-based performance assessment.

Part 1

Introduction

1 Key Themes, Constructs and Interdisciplinary Perspectives in Second Language Pronunciation Assessment

Talia Isaacs and Pavel Trofimovich

Assessment of Second Language Pronunciation: Where We Are Now

After a period of relative neglect, second language (L2) pronunciation has experienced a resurgence of interest among applied linguistics researchers and L2 practitioners, with several indicators signalling growing momentum. For example, the past decade has witnessed the emergence of pronunciation-specific special journal issues (e.g. Cardoso & Trofimovich, 2014), invited symposia (e.g. Derwing & Munro, 2010), webinars and Electronic Village Online sessions organized by the pronunciation special interest group of professional teaching associations (e.g. Harding & Selman, 2014), research timelines (e.g. Munro & Derwing, 2011), meta-analyses (e.g. Lee *et al.*, 2015), and encyclopaedia volumes or handbooks (Reed & Levis, 2015). In addition, evidence of the growing interest in L2 pronunciation research is reflected in the establishment of the annual *Pronunciation in Second Language Learning and Teaching* (PSLLT) conference and proceedings in 2009 and, more recently, in the launch of the *Journal of Second Language Pronunciation* in 2015 – a symbol of the professionalization of the field. These developments have been accompanied by a substantial overall increase in the proportion of pronunciation-relevant articles published in applied linguistics journals over the past few years (Levis, 2015), which is key to the reintegration of pronunciation research into the applied linguistics research mainstream after decades of being sidelined. Several recent graduates with pronunciation expertise have also launched into academic positions at

international universities and are, in turn, training a new generation of pronunciation proponents, assuring L2 pronunciation a bright future in research and teacher training in the years to come, although there is much more work to be done (Derwing & Munro, 2015).

Pronunciation is, by its nature, interdisciplinary, drawing on research traditions in psycholinguistic, sociolinguistic and speech sciences and strongly interfacing with work in second language acquisition (SLA) and L2 pedagogy. There have been developments in all of these areas, although few common platforms for discussion exist, as the scholarly discourse, methodologies and research priorities vary substantially across domains. Notably, much of the renewed applied pronunciation related activity over the past several decades has been conducted by SLA researchers and research practitioners interested in teacher training and, to a lesser extent, by those researching the use of an L2 as a lingua franca across the globe. Interest in L2 pronunciation from within the language assessment community specifically, which includes both researchers and practitioners (e.g. exam board staff), has taken much longer to ignite. For example, there is no dedicated book on assessing L2 pronunciation in the foundational *Cambridge Language Assessment* series to accompany books on assessing other language components (e.g. grammar and vocabulary, although assessing pragmatics is similarly not featured). Pronunciation also plays only a peripheral role in books on assessing L2 speaking (Fulcher, 2003; Luoma, 2004) and was singled out as not having been included in Fulcher's (2015) research timeline on the topic. Until recently, there has also been little acknowledgement of the absence of pronunciation from the L2 assessment research agendas (Isaacs & Thomson, 2013), or of its often peripheral role in assessing L2 speaking proficiency, including in scales, where it has either been unmodelled or inadequately operationalized (Harding, 2013, this volume; Isaacs *et al.*, 2015).

The 2011 *Language Testing Research Colloquium* marked the 50th anniversary of the publication of Lado's (1961) seminal book, *Language Testing*, which is widely considered to signify the birth of the language assessment field (Spolsky, 1995). Over half a century later, Lado's work remains the only non-thesis single-authored book-length treatment on pronunciation assessment (among other topics) and, hence, the existing authority on designing and administering pronunciation tests, despite some key concepts being out of date (Isaacs, 2014). However, there are recent indications that pronunciation assessment is emerging from its time warp. For example, whereas only two pronunciation-focused articles were published in the longest standing language assessment journal, *Language Testing*, in its first 25 years of publication (1984–2009; Isaacs, 2013), seven articles have appeared in the five-year period since (2010–2015; Levis, 2015). Pronunciation assessment has also been featured in major events targeting the L2 speaking construct (e.g. the 2013 *Cambridge Centenary Speaking Symposium*) and in at least four externally funded TOEFL and IELTS research projects since 2010, a

topic hitherto rarely focused on in the validation of high-stakes tests. This implies that pronunciation is increasingly being viewed as integral to the L2 speaking construct.

Beyond the piecemeal contributions of individual researchers, a more sustained shift in attention back to pronunciation from the language assessment community at large has been seen in the introduction of fully automated standardized L2 speaking tests (e.g. Pearson's Versant test and Educational Testing Services' SpeechRater), which place considerable weight on acoustic and temporal measures in scoring (Kang & Pickering, 2014; Zechner *et al.*, 2009). The launch of fully automated tests in the international language testing market (e.g. the Pearson Test of English Academic for university entrance purposes) fed into a rigorous field-wide debate on machine-mediated automated scoring in the first decade of the 21st century (e.g. Chun, 2006, 2008; Downey *et al.*, 2008), which has arguably evolved into more pragmatic acceptance of the inevitability of the use of automated speech recognition technology during the second decade (e.g. Isaacs, 2016; Xi, 2010, 2012).

The growing use of English as a lingua franca in diverse international settings brought about by economic globalization and technological advancements has catapulted the issue of defining an appropriate pronunciation standard in L2 listening and speaking tests (e.g. Canagarajah, 2006; Elder & Davies, 2006; Jenkins, 2006; Ockey & French, 2014), in light of growing attention to proposals for supplementing (if not supplanting) the native speaker standard. Such discussions are permeating the decades-long language testing literature on international teaching assistants (ITA), with pronunciation-relevant research strands now focusing on identifying the linguistic features that are most important for being intelligible or easily understood by listeners, in addition to identifying sources of listener bias (e.g. listener background characteristics, such as differential exposure to particular varieties of L2 accented speech) that could have bearing on their judgements of oral performance, instructional competence or other social measures (e.g. Hsieh, 2011; Isaacs, 2008; Kang, 2008, 2012).

Bringing Together Different Research Strands

Although there are signs of growing interest in L2 pronunciation assessment among researchers and educational practitioners, there is, as yet, no synthesis of work beyond single book chapters in edited volumes that tend to target either audiences of primarily language testers (e.g. Isaacs, 2014), or predominantly SLA-oriented pronunciation researchers (e.g. Levis, 2006), with little apparent crossover between these communities. Consolidating knowledge on pronunciation assessment is sorely needed to keep pace with current advancements, promote a baseline level of understanding of relevant issues, spearhead interdisciplinary dialogue, guide teaching and test

development, and inform future research directions. This volume seeks to fill this gap by bringing to light the insights from assessing other skills (e.g. listening, writing) in addition to drawing on perspectives from research in speech sciences, SLA, psycholinguistics and sociolinguistics, including lingua franca communication, with concrete implications for pronunciation assessment. This edited collection thus pools the expertise of authors from different research communities to establish a common platform by which to carry issues forward in a research area that is increasingly assuming a higher profile and gaining currency in all domains within applied linguistics.

The edited collection caters to a mixed audience of L2 researchers, graduate students, teacher-educators and exam board staff with varying levels of expertise in pronunciation and assessment. It is conceived of as the first point of reference for readers from different disciplinary backgrounds, bringing to the fore topical issues and challenges that relate to formal and informal assessments of L2 pronunciation in classroom, research and real-world contexts. The edited volume is thus likely to be informative to both a new generation of researchers hoping to make inroads in pronunciation and/or assessment, and experienced pronunciation researchers who wish to consult and cite high-calibre work both within and beyond their specific areas of expertise. Although not explicitly tackling problems to do with developing and validating L2 pronunciation tests (e.g. item writing), which remains a tangible gap in the literature (Isaacs, 2014), the concrete implications for pronunciation assessment in each study are likely to address at least some important conceptual and practical issues and to generate further thought and discussion. Due to its interdisciplinary nature, the edited volume is likely to cater to students, researchers and practitioners with wide-ranging interests in applied linguistics that extend beyond pronunciation.

The chapters, which together span the methodological spectrum (quantitative, qualitative, mixed methods), represent the breadth of research traditions used to examine the linguistic and non-linguistic phenomena relevant to L2 pronunciation assessment. The chapters also include state-of-the-art reviews, empirically grounded contributions and research commentaries that interface with different aspects of pronunciation and assessment, elucidating key issues and underscoring implications for further research and practice. Despite the substantive and methodological breadth of each contribution making up the collection, the following principles apply to all:

(1) Each chapter is written in clear and accessible language for an audience of academics, graduate students and L2 teaching and testing professionals with varying expertise in L2 pronunciation and assessment.
(2) Key definitions of relevant terms are provided within the context of each chapter to promote an understanding of the definition of major constructs for the purposes of the reported study.

(3) In the case of research reporting, justification for key methodological decisions is provided to render the rationale behind novel procedures or adherence to research conventions more transparent.
(4) Each chapter concludes with a section delineating concrete implications for research and practice in L2 pronunciation assessment or future directions.

Structure of the Book

This book consists of 14 chapters, which can be read in sequence or as stand-alone units, featured in four main sections.

Part 1: Introduction

The chapters in this part, including the state-of-the-art overview in this introductory chapter, cover fundamental concepts in L2 pronunciation research, centring on ways in which major constructs are defined and operationalized, including problematizing pronunciation assessment instruments used by human raters.

Part 2: Insights from Assessing Other Language Skills and Components

This part focuses on the learning and assessment of other L2 skills and components, with chapters on assessing fluency, writing and listening. The assessment of these areas of ability has been more extensively researched than the assessment of L2 pronunciation, and insights could be useful in informing the future development of the field.

Part 3: Perspectives on Pronunciation Assessment From Psycholinguistics and Speech Sciences

This part consists of empirical studies grounded in research in psycholinguistic and speech sciences, including work on individual differences in listener (rater) characteristics and different objective and subjective ways of measuring the linguistic properties of L2 speech.

Part 4: Sociolinguistic, Cross-cultural and Lingua Franca Perspectives in Pronunciation Assessment

This part focuses on the implications and applications of pronunciation teaching and assessment in various cultural, educational and lingua franca contexts, including the role of the native speaker as an assessment standard. These contributions provide a unique perspective to the volume by

contextualizing pronunciation assessment within the complexities of present-day multilingual, cross-cultural and educational spaces.

Part 5: Concluding Remarks

The concluding part summarizes, synthesizes and discusses the nature of the innovation of each contribution and of the volume as a whole. It concludes with future directions for L2 pronunciation research and practice formulated as research themes and questions that are likely to be the subject of further investigation.

Key Concepts and Definitions

As stated above, this book responds to the urgent need to consolidate current expertise on L2 pronunciation assessment by bringing together insights and highlighting pedagogical and assessment implications from within the applied linguistics community that are of relevance to language assessment researchers and practitioners on a common platform. Having a single forum for bringing together different voices is a preliminary means of arriving at a common understanding of key issues, understanding the breadth of approaches, and charting future directions from an informed and interdisciplinary perspective, which is the overarching goal of this volume.

It seems fitting that a book that includes contributions from members of different research communities would begin by establishing common threads and providing definitions as a means of synchronizing across the different areas. However, without wanting to impose a priori definitions to the authors when approaching them for contributions – because they are conducting work on different facets of L2 pronunciation and/or assessment with distinct areas of expertise – providing common definitions of at least some key terms, particularly in L2 pronunciation, for the benefit of end-users is, at this stage, unfeasible. This is because there has been little cross-talk across fields and no precedent in terms of edited volumes on L2 pronunciation assessment that encompass the breadth of the research and practical applications presented in this collection. Although contributors are, in some cases, writing about similar issues, they tend to be approaching problems from different perspectives and, for the most part, speaking in very different languages, with different underlying assumptions and understanding of key issues, which they strive to clarify with transparency through the course of their chapter, and with discernibly different research priorities. Thus, the task of providing all-encompassing definitions that pervade all of the contributions – for example, for a term such as 'intelligibility', which has been defined and measured in numerous ways in the literature (Isaacs, 2008) in a similar way to a term such as 'fluency' (Koponen & Riggenbach, 2000) is

difficult, with different shades of meaning coming to light in different chapters as the concept is discussed in both broader conceptual terms, and in narrower operational meanings in the context of individual studies.

Despite these challenges, we feel that it is both possible and appropriate to clearly define the terms 'pronunciation' and 'assessment' that appear in the title. 'Pronunciation', in the way it was conceived of for this edited volume, encompasses (1) individual consonant and vowel sounds, commonly referred to in the literature as 'segments', and (2) features that span a larger unit than a single segment, such as word stress, rhythm and intonation, referred to synonymously in the literature as 'suprasegmentals' or 'prosody' – terms that are, therefore, used interchangeably in this volume. However, the reader should be aware that language tests, including rating scales, may have their own operational definitions of these terms that diverge from these meanings (Isaacs *et al.*, 2015).

Following Bachman's (2004) expanded view of assessment, the term 'assessment' in this volume broadly refers to the process of information gathering (e.g. about an L2 learner's or test taker's ability), potentially from multiple and varied sources on the variable(s) of interest, including generating information about what learners can do to feed into the teaching cycle. In contrast a 'test' refers more specifically to a particular type of assessment involving the elicitation of an L2 learner's or test taker's performance followed by inferences or decision making on the basis of that performance, generally informed by a test score or a numerical indicator from a score report. Therefore, all tests are also assessments, whereas not all assessments are tests, although tests are a very common and, due to their often high stakes, the most high-profile form of assessment.

We hope that this volume will be viewed as a trendsetter in a burgeoning field that is steadily gaining momentum, consolidating knowledge on current practice across disciplinary areas and driving the conversation forward. We also hope that it will help establish commonalities across research areas and facilitate greater consensus and agreement about key issues, terminology and best practice in L2 pronunciation research and assessment moving forward.

References

Bachman, L.F. (2004) *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Canagarajah, S. (2006) Changing communicative needs, revised assessment objectives: Testing English as an International Language. *Language Assessment Quarterly* 3 (3), 229–242.

Cardoso, W. and Trofimovich, P. (2014) Second language speech perception and production: Implications for language teaching. Special Issue of *Canadian Modern Language Review* 70 (4).

Chun, C.W. (2006) An analysis of a language test for employment: The authenticity of the PhonePass test. *Language Assessment Quarterly* 3 (3), 295–306.

Chun, C.W. (2008) Comments on 'evaluation of the usefulness of the Versant for English test: A response': The author responds. *Language Assessment Quarterly* 5 (2), 168–172.
Derwing, T.M. and Munro, M.J. (2010) Symposium – accentuating the positive: Directions in pronunciation research. *Language Teaching* 43 (3), 366–368.
Derwing, T.M. and Munro, M.J. (2015) *Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research*. Amsterdam: John Benjamins.
Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M. and Van Moere, A. (2008) Evaluation of the usefulness of the Versant for English test: A response. *Language Assessment Quarterly* 5 (2), 160–167.
Elder, C. and Davies, A. (2006) Assessing English as a lingua franca. *Annual Review of Applied Linguistics* 26 (1), 232–301.
Fulcher, G. (2003) *Testing Second Language Speaking*. London: Pearson.
Fulcher, G. (2015) Assessing second language speaking. *Language Teaching* 48 (2), 198–216.
Harding, L. (2013) Pronunciation assessment. In C.A. Chapelle (ed.) *The Encyclopedia of Applied Linguistics*. Hoboken, NJ: Wiley-Blackwell.
Harding, L. and Selman, A. (2014) Report on the joint IATEFL PronSIG/TESOL SPLIS group webinar. *Speak Out!* 51, 63–69.
Hsieh, C.-N. (2011) Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency. *Spaan Fellow Working Papers in Second or Foreign Language Assessment* 9, 47–74.
Isaacs, T. (2008) Towards defining a valid assessment criterion of pronunciation proficiency in non-native English speaking graduate students. *Canadian Modern Language Review* 64 (4), 555–580.
Isaacs, T. (2013) Pronunciation. In *Cambridge English Centenary Symposium on Speaking Assessment* (pp. 13–15). Cambridge: Cambridge English Language Assessment.
Isaacs, T. (2014) Assessing pronunciation. In A.J. Kunnan (ed.) *The Companion to Language Assessment* (pp. 140–155). Hoboken, NJ: Wiley-Blackwell.
Isaacs, T. (2016) Assessing speaking. In D. Tsagari and J. Banerjee (eds) *Handbook of Second Language Assessment* (pp. 131–146). Berlin: DeGruyter Mouton.
Isaacs, T. and Thomson, R.I. (2013) Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly* 10 (2), 135–159.
Isaacs, T., Trofimovich, P., Yu, G. and Chereau, B.M. (2015) Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS Research Reports Online* 4, 1–48.
Jenkins, J. (2006) The spread of EIL: A testing time for testers. *ELT Journal* 60 (1), 42–50.
Kang, O. (2008) Ratings of L2 oral performance in English: Relative impact of rater characteristics and acoustic measures of accentedness. *Spaan Fellow Working Papers in Second or Foreign Language Assessment* 6 (1), 181–205.
Kang, O. (2012) Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly* 9 (3), 249–269.
Kang, O. and Pickering, L. (2014) Using acoustic and temporal analysis for assessing speaking. In A.J. Kunnan (ed.) *The Companion to Language Assessment* (pp. 1047–1062). Hoboken, NJ: Wiley-Blackwell.
Koponen, M. and Riggenbach, H. (2000) Overview: Varying perspectives on fluency. In H. Riggenbach (ed.) *Perspectives on Fluency* (pp. 5–24). Ann Arbor, MI: University of Michigan Press.
Lado, R. (1961) *Language Testing: The Construction and Use of Foreign Language Tests*. London: Longman.

Lee, J., Jang, J. and Plonsky, L. (2015) The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics* 36 (3), 345–366.

Levis, J.M. (2006) Pronunciation and the assessment of spoken language. In R. Hughes (ed.) *Spoken English, TESOL and Applied Linguistics: Challenges for Theory and Practice* (pp. 245–270). New York: Palgrave Macmillan.

Levis, J. (2015) Pronunciation trends across journals and the Journal of Second Language Pronunciation. *Journal of Second Language Pronunciation* 2 (1), 129–134.

Luoma, S. (2004) *Assessing Speaking*. Cambridge: Cambridge University Press.

Munro, M.J. and Derwing, T.M. (2011) Research timeline: Accent and intelligibility in pronunciation research. *Language Teaching* 43 (3), 316–327.

Ockey, G. and French, R. (2014) From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*. Published online. doi:10.1093/applin/amu060.

Reed, M. and Levis, J. (eds) (2015) *The Handbook of English Pronunciation*. Malden, MA: Wiley Blackwell.

Spolsky, B. (1995) *Measured Words: The Development of Objective Language Testing*. Oxford: Oxford University Press.

Xi, X. (2010) Automated scoring and feedback systems for language assessment and learning. Special Issue of *Language Testing* 27 (3).

Xi, X. (2012) Validity and the automated scoring of performance tests. In G. Fulcher and F. Davidson (eds) *The Routledge Handbook of Language Testing* (pp. 438–451). Abingdon: Routledge.

Zechner, K., Higgins, D., Xi, X. and Williamson, D.M. (2009) Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication* 51 (10), 883–895.

2 What Do Raters Need in a Pronunciation Scale? The User's View

Luke Harding

Introduction

Pronunciation scales have been shown to be highly problematic to design and implement, with descriptors suffering from inconsistencies, vague language, conflated constructs and unclear trajectories (see Harding, 2013; Isaacs, 2013). In addition, pronunciation presents an area of judgement where a rater might rely heavily on his or her own conceptualization of the construct when a scale becomes difficult to use (Harding, forthcoming). It is therefore essential that users' experiences in applying pronunciation scales are understood and considered in the scale design process. This chapter will present a study that investigated the usability of a particular pronunciation scale, with a view to abstracting from raters' experiences to a set of general principles for the future design of usable pronunciation rating instruments.

Background

A rationale for scale usability research

Designing a rating scale is a challenging task for any prospective language test developer. The difficulty comes in two parts: (1) knowing what information to include in scale descriptors at different levels; and (2) ensuring that the scale will be interpreted correctly and consistently by raters. Much of the literature on rating scales and rater behaviour has been framed around these two challenges. On the first point – knowing what information to include in descriptors – a now common set of methods is recommended for scale development depending on a designer's level of expertise and the resources at

hand. Fulcher (2003) describes the two broad scale design approaches as 'intuitive' and 'empirical'. The former relies on procedures such as expert or committee judgement to develop a set of criteria, perhaps modifying an existing scale. The latter consists of methods such as scaling descriptors (the procedure undertaken to create the Common European Framework of Reference (CEFR)), empirically derived binary-choice boundary definition scales (EBBs), or data-driven scale construction where performance data is drawn on to identify criterial features across levels. Empirical methods are perceived as the more rigorous approach, and accounts of empirical scale development can be found in the research literature: for example, North (2000) for scaling methods, Upshur and Turner (1995) for EBBs, and Fulcher (1996) for the data-driven approach (see also Knoch, this volume).

On the second point – ensuring that the scale will be interpreted correctly by raters – research has also explored what happens after a scale has been developed and is in use, focusing on whether raters interpret scale descriptors in construct-relevant ways, identifying factors which guide decision-making processes in assigning grades, and attempting to understand the bases of divergent rating behaviour; this is the growing field of rater cognition (see Bejar, 2012). Researchers have considered rater decision making in relation to writing (Baker, 2012; Cumming *et al.*, 2002), speaking (May, 2009; Orr, 2002), and the grading of limited production tasks in listening assessment (Harding *et al.*, 2011). One key finding of studies of this kind is that scales themselves have a limited capacity for ensuring valid interpretation and consistent application among raters. As Bejar (2012: 4) notes, 'Important as rubrics are, it is still a fairly abstract document from the point of view of the scorers'. In recognition of this, ongoing rater training and other supporting documentation such as benchmarked performances are often recommended to scaffold the role of the scale in the rating process.

There is, however, a point at the nexus of scale design and scale use that has been less explored in the research literature, and this concerns the usability of the rating scale. Usability is here defined simply as 'ease of use' – a definition that has its roots in the field of software development and human–computer interaction, and is associated with the paradigm of user-centred design (Norman, 2002), an approach that places the needs and capabilities of the user as the primary consideration throughout the design process. A focus on usability is warranted in light of the critiques that have been made of many rating scales in the research literature. Apart from the reductionism that is a necessary feature of any rating scale (see Van Moere, 2013), scales may suffer from design problems that make them less useful tools for their ultimate users – raters. Scales may be overly complex (Fulcher, 1996), overly simplistic (Cumming *et al.*, 2002), or fail to encode the features of performances most salient to judges (Mcnamara, 1996). When a scale has been intuitively or empirically derived, or if a rater is not able to work easily with the scale because it has not been designed with the rater's needs in

mind, then the rater's adherence to the scale may be reduced (see also Bejar, 2012). There is value, then, in seeking the views of raters themselves on the qualities of scales that affect usability, and in using these data to feed back into the design process of specific instruments. There is also the potential for usability studies to inform general principles about scale design across specific skills which may be helpful for other practitioners.

Usability in rating scale development research

There are numerous examples in the research literature where rating scales have been criticized from an assessment expert's perspective (e.g. Brindley, 1998; Horner, 2013; Turner & Upshur, 2002). Capturing the rater's perspective *in situ*, however, is less common. This is not to say that field testing of scales with raters is not done regularly in practice; in fact, it may be considered a key step in the scale development or revision cycle, particularly in large-scale testing programmes. However, the data yielded in rater consultation sessions are not often reported as the subject of research. One exception is a recent article by Galaczi *et al.* (2011a) on the revision of speaking scales for Cambridge English exams. Galaczi *et al.* describe a verbal report study in which experienced raters commented on their experience of applying a set of revised scales. Raters approved of new positively worded descriptors, and commended a shift in the wording of descriptors towards greater specificity, avoiding terms like 'mainly', 'may', 'might' and 'usually' (Galaczi *et al.*, 2011a: 229). However, there were still criticisms of other ambiguous terms such as 'good degree of control' (Galaczi *et al.*, 2011a: 229). The raters' comments were fed back into the development process, and also informed ongoing rater training and moderation. For the most part, though, the scales were accepted by the raters, and difficulties did not extend beyond terminological problems.

Where they exist, the most extensive treatments of scale usability have been provided in the literature on writing assessment. Knoch (2009), for example, sought raters' comments on two rating scales for diagnostic writing – one that was in current use, and one that was newly developed. Raters commented on several aspects related to usability across both scales, including: the explicitness/vagueness of descriptors; the difficulty of distinguishing between scale categories; and instances where the scale did not reflect those elements of the construct salient to raters. A similar process of collecting rater feedback was undertaken by Harsch and Martin (2012), although in their case the consultation with raters occurred concurrently with a rater training procedure. Findings from their study showed, again, that vague wording was problematic for raters (e.g. 'repertoire' and 'range'), and that terms expressing possibility or probability (e.g. 'may show') were difficult for raters to apply (reflecting the findings of Galaczi *et al.*, 2011a). Raters also found it difficult to interpret descriptors relating to control or accuracy alongside more positively worded statements, leading Harsch and Martin to

conclude that different descriptor wordings may lead to 'cognitively different task[s]' for raters in applying those descriptors.

Turning specifically to pronunciation, Yates *et al.*'s (2011) study of examiner perceptions of the revised IELTS pronunciation scale represents the only detailed examination to date of scale usability focusing on this particular skill. In their study, a group of raters was invited to comment on a revised pronunciation scale before and after use via questionnaires. A separate group of raters provided comments on difficulties in applying the scale through verbal report protocols. The study was designed to investigate some key changes to the IELTS pronunciation scale, one of the most notable being the shift from four-band levels (with level descriptors for bands 2, 4, 6 and 8) to a nine-band scale in line with the three other criteria for IELTS speaking (fluency and coherence, lexical resource, and grammatical range and accuracy). Critiques of the new scale related to two broad areas: (1) the wording of descriptors at Bands 3, 5 and 7; and (2) the overlap between pronunciation and other speaking skills. On the first point, like those studies mentioned above, a lack of explicitness in terminology was a key source of difficulty for raters in applying the levels. However, most of these comments focused on the descriptors for the new bands – 3, 5 and 7 – which require an estimation of the extent to which candidates have met a previous level's descriptor and what they are able to do at the level immediately above. For example, Band 5 includes the descriptor: 'all the positive features of Band 4 and some, but not all, of the positive features of Band 6' (IELTS, 2015; see also Yates *et al.*, 2011). Raters problematized the calculations involved, particularly as some descriptors at neighbouring levels were expressed in negative terms, a finding that has also recently been echoed in Isaacs *et al.* (2015). There was also a sense that the lack of clear descriptors at these levels was a 'cop out' (Yates *et al.*, 2011: 30). On the second point, raters commented that there were overlaps between the pronunciation scale and the fluency and coherence scale, with elements such as speech rate, repetition and rhythm influencing judgements on both criteria. Recommendations of the study included developing clearer descriptors for Bands 3, 5 and 7, and reconsidering the distinction between the pronunciation scale and the fluency and coherence scale.

An instrumental case: The CEFR Phonological control scale

While the studies above provide insight into some general principles of scale usability – particularly with respect to the wording of descriptors – there remains a need for further exploration of this issue. First, much of the research to date has focused on the use of writing scales (see Knoch, this volume), where the processes of rating will necessarily follow different patterns from the real-time rating of speech. Secondly, although the usability of the IELTS scale has been explored in some detail by Yates *et al.* (2011), the raters were commenting on a 'polished' scale which had already been

introduced, and which had been the subject of in-house development procedures. For this reason, the lessons that can be learned from the findings are limited and, to a certain extent, specific to that scale.

Therefore, for the purposes of this study, the CEFR Phonological control scale was chosen to represent an 'instrumental case' (see Stake, 1995) for the purposes of exploring the nature of pronunciation scale usability. The CEFR is a set of guidelines containing descriptions of language proficiency across six levels (A1, A2, B1, B2, C1 and C2). It was produced by the Council of Europe in the 1990s, and over the past decade has come to have significant impact on language learning, teaching and assessment globally. The Framework (see Council of Europe, 2001) consists of sets of scales including a global scale, and illustrative scales for a range of different communicative activities (e.g. reading, writing, listening and speaking). While there are no explicit references to pronunciation in the CEFR global scale descriptors, the Phonological control scale – one of the illustrative 'linguistic' scales – provides criterial descriptions of pronunciation ability over the first five levels of the CEFR: A1–C1 (there is currently no descriptor for C2). The full scale is presented in Table 2.1.

The Phonological control scale is ostensibly a 'user-oriented' scale (i.e. having a reporting function) rather than an 'assessor-oriented' scale (i.e. 'guiding the rating process'; see Alderson, 1991: 73); however, as with other CEFR scales it has, in practice, served as the basis for rating scale development (e.g. Pearson Test of English General; see Pearson Education, 2012). Raters' comments on the scale may therefore have useful applications for other pronunciation scale development projects based on the CEFR. More importantly, because the scale is currently thought to be under-specified (see Galaczi *et al.*, 2011b), and has been critiqued by researchers as lacking consistency, explicitness and a clear underlying construct (Harding, 2013, forthcoming; Horner,

Table 2.1 The CEFR Phonological control scale

C2	As C1
C1	Can vary intonation and place sentence stress correctly in order to express finer shades of meaning.
B2	Has acquired a clear, natural, pronunciation and intonation.
B1	Pronunciation is clearly intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur.
A2	Pronunciation is generally clear enough to be understood despite a noticeable foreign accent, but conversational partners will need to ask for repetition from time to time.
A1	Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by native speakers used to dealing with speakers of his/her language group.

Source: © Council of Europe (2001: 117).

2013; Isaacs & Trofimovich, 2012), the CEFR phonological control scale was considered a useful instrument for eliciting raters' comments on a range of usability problems. As previous research has often focused on scales that are in the final stages of development or revision, where raters' input was used for fine-tuning, it would be illuminating to see the full range of responses as raters deal with this scale which has been identified among pronunciation and/or assessment researchers as needing improvement.

Aim and Research Questions

The aim of this study, then, was to attempt to explore raters' experiences in using a rating scale for pronunciation with a view to establishing the needs and preferences of raters in future pronunciation scale design. The specific research questions the study set out to answer were:

(1) What aspects of the CEFR Phonological control scale, if any, do raters problematize?
(2) What inferences can be drawn from raters' identified problems for the design of pronunciation scales generally?

Methodology

Research context

The data for this study were drawn from a larger mixed-methods research project which had the principle aim of investigating the construct underlying the CEFR Phonological control scale, specifically its orientation towards either a nativeness principle (where the ultimate target is nativelike pronunciation) or a comprehensibility principle (where the target is ease of understanding) (reported in Harding, forthcoming; see also Levis, 2005). The current study explores data elicited specifically from a focus group that took place during this project, where numerous issues related to the usability of the CEFR phonological control scale were raised that were beyond the scope of the larger project.

Focus group methodology

Participants were invited to join a focus group to discuss their experiences working with the CEFR phonological control scale (see 'Data collection procedures' section for detailed procedures). Focus groups are a less common methodology in language testing research (although see Harding *et al.*, 2011; Isaacs *et al.*, 2011, 2015; Ryan, 2007). However, they present several advantages over traditional one-on-one interview methods; they are efficient, allow

Table 2.2 Summary of raters' experience (Harding, forthcoming)

Rater ID	*Summary of experience*
R1	EAP tutor (3 years)
R2	ESOL/EFL teacher (several years); regularly assesses pronunciation one-on-one
R3	Trinity College London examiner (7 years); EAP/EFL teacher (12+ years)
R4	Trinity College London examiner (4 years); EFL/EAP teacher (36 years)
R5	IELTS examiner (14 months); EFL teacher (7 years)
R6	IELTS and Cambridge Main Suite examiner (1 year); EFL teacher (10+ years)
R7	IELTS examiner (12 years); Trinity College London examiner (4 years)
R8	French teacher/lecturer; regularly conducts oral assessments
R9	FCE and IELTS examiner (20 years); IELTS examiner trainer (5 years)

Source: Harding (forthcoming).
Notes: EAP = English for academic purposes; ESOL = English for speakers of other languages; EFL = English as a foreign language; IELTS = International English Language Testing System; FCE = First Certificate English.

for interaction between participants, and can be less researcher-directed. In the context of this study, it is also useful to note that focus groups are a common method of collecting early-stage usability data in other fields.

Participants

Nine experienced raters were invited to participate in the study. The recruitment process purposefully targeted raters who had reasonably high levels of experience in assessing language in the classroom or in more formal examining contexts. While there was some variability in the level of experience among the final group, this variability represented the different levels of expertise one might expect in any type of rater cohort, thus enhancing the ecological validity of the study. Details on individual raters' experience are provided in Table 2.2.

All raters were female. The group was a mix of native speakers of English (six) and highly proficient/bilingual users of English (three), covering Sinhalese, French and Slovene.

Data collection procedures

Prior to taking part in the focus group, raters completed two initial stages:

(1) Raters worked through a set of familiarization activities at home. This ensured that all participants had a working knowledge of the CEFR phonological control scale descriptors prior to the rating session. The activities were modelled on the familiarization tasks recommended as

the first step in standard setting (see Council of Europe, 2009) and involved studying the scale and matching descriptors to levels as a form of self-assessment.

(2) Raters then attended a one-day rating session at Lancaster University, the first stage of which involved using the Phonological control scale to rate 44 speech samples produced by second language (L2) users of English. The speech samples represented a range of proficiency levels and first language (L1) backgrounds, although with a sizable proportion of Chinese L1 background speakers ($n = 30$). Each speech sample was an extemporaneous narrative based on a common six-panel picture description task (drawn from Heaton, 1975). Speech samples were between 55 and 70 seconds long. The rating process took around 90 minutes including breaks.

Following the initial stages – familiarization and rating – all nine raters then took part in the focus group discussion. The session took place in a meeting room immediately following the rating stage. Raters were able to refer to their rating sheets, where they had also taken notes while rating. The discussion was audio-recorded, with the researcher acting as the focus group moderator. Three broad questions guided the focus group discussion:

(1) Did you find the CEFR scale descriptors easy or difficult to apply? Why?
(2) Were there any speakers who you couldn't place on the scale? Who were they? Why?
(3) Do you think the scale captured the range of pronunciation you heard effectively?

Apart from providing probes and occasionally steering topics back towards the main themes, the moderator played a minimal role for much of the discussion. The focus group came to a natural conclusion after approximately 45 minutes.

Analysis

The focus group was first transcribed for content. The analysis then involved the identification of thematic units related to the research questions. The process of thematic analysis followed the procedures recommended by Braun and Clarke (2006: 87–93), which include (in summary):

(1) familiarizing yourself with the data through transcribing, reading and re-reading the data;
(2) generating initial codes across the dataset;
(3) searching for potential themes by collating codes;
(4) reviewing themes to ensure they make sense with relation to the codes and the dataset as a whole;

(5) defining and naming themes;
(6) producing the report.

The details of themes that emerged from this analysis are described in the next section.

Findings

Following the thematic analysis of raters' discussions in the focus group, four macro themes were identified in relation to the usability of the Phonological control scale:

(1) clarity;
(2) conciseness;
(3) intuitiveness;
(4) theoretical currency.

These themes, and their related sub-themes, are discussed in turn with illustrative examples from the dataset provided in the sections below.

Clarity

The theme of 'Clarity' captures comments which related to the overall comprehensibility (ease of understanding) of the scale. Raters referred to three different problems in relation to clarity: (1) the coherence of the scale; (2) specific terminological problems that created confusion; and (3) descriptors that appeared to be irrelevant to assessing pronunciation.

Scale coherence

The most commented upon aspect of the scale was its coherence – the extent to which the scale was structured logically. Specifically, the erratic appearance of particular elements of pronunciation across levels was a key topic in the focus group. One example concerns the place of intonation in the scale, which appears at the B2 and C1 levels, but is not mentioned at any other levels of the scale. Excerpt 2.1 illustrates that raters were unsure about how the absence of intonation-related descriptors at the lower levels should be interpreted.

Excerpt 2.1

Rater 7: This is what I mean about inconsistency … they don't mention … you know intonation only occurs in B2, why isn't it mentioned throughout?

Rater 2: Because it's part of accent. Intonation is part of an accent, you've got articulation, intonation, stress all those different things are all …
Rater 7: But I think it's almost like intonation doesn't particularly matter until you get to B2.
Rater 3: It does seem that way.

This exchange suggests that there was an expectation that integral elements of pronunciation – such as suprasegmental features – should be referred to across the entire scale.

A second coherence problem identified by the raters was the appearance of descriptors for foreign accent at A2 and B1 levels, with no further reference made to foreign accent at the higher levels. This feature of the scale was the focus of intense early discussion in the focus group. A representative example is shown in Excerpt 2.2.

Excerpt 2.2

Rater 7: … I'm not very happy with this term 'foreign accent' which is used in B1 and A2, um, and doesn't then appear afterwards which strikes me as a bit odd because, you know, all of them had a foreign accent so … if you were to apply these strictly you wouldn't want to give anybody above a B1. So I think that's a flaw within these descriptors.

The raters reported overcoming this problem by balancing the presence of accent with the comprehensibility of the speaker, allowing them to award B2 and C1 levels even if speakers had perceptible foreign accents (see Excerpt 2.3).

Excerpt 2.3

Rater 8: It's to me an accent is an accent, we all have it but does it actually prevent, does it …
Rater 3: Impede.
Rater 8: Impede the message …

Excerpt 2.3 suggests that where the scale was found to lack logic, the raters drew on their own strategies for making sense of the descriptors. Harding (forthcoming) found that the strategies raters used in dealing with this particular coherence problem were probably not uniform – a fact that only serves to underline the need for coherence to be a target quality of the scale's descriptors.

It is important to note that North (2014) has recently defended this particular type of criticism of the CEFR scales more generally, pointing out that

the CEFR was developed using a 'salient features' approach. Under this approach certain elements of a given scale may appear sporadically: '[f]eatures land where they were calibrated on the scale. There is no attempt to describe everything at every level' (North, 2014: 27). The sound theoretical basis of this measurement approach, however, does not detract from the fact that the raters here appeared to view representation of key dimensions of pronunciation across the whole scale as a central element of usability. Whether this requirement is a vestige of what the raters are accustomed to from prior rating experience, or the expression of a basic need to understand how features develop over levels, remains uncertain.

Terminological problems

A second criticism of the clarity of the scale was the use of fuzzy or ambiguous terminology. This finding was anticipated, as the previous literature on scale usability (e.g. Galaczi *et al.*, 2011a) and on scale design in general (e.g. Fulcher, 2003; Van Moere, 2013) have problematized the vague terminology commonly employed in rating scales. Within the focus group, raters identified the term 'natural', which appears at the B2 level in reference to intonation, as particularly problematic.

Excerpt 2.4

Rater 6: Like the word 'natural' now in B2 it's 'natural' intonation and in C1 it's 'varying' intonation then what, what do you really mean by natural?

Rater 5: Mm, natural is a difficult word.

'Natural' in the context of the scale is only interpretable with relation to some external standard of 'naturalness' in intonation. At the same time, the presence of the more concrete notion 'can vary intonation' at a higher level suggests that variation is a feature beyond what would be expected in a 'natural' intonation. The trajectory of the scale, therefore, asserts its own definition of 'naturalness', which appears to be at odds with the raters' own conceptualizations of what 'naturalness' might entail, as evidenced in the rhetorical question, 'what do you really mean by natural?'.

Another difficult term for raters to deal with was 'noticeable', which is used in the descriptor phrase 'noticeable foreign accent' at A2. From the excerpt below, it is apparent that the term 'noticeable' does not make sense in relation to the neighbouring B1 descriptor where foreign accent is described as 'sometimes evident'.

Excerpt 2.5

Rater 3: But I think the descriptors are woolly. I, sorry but if I look at B1 where it says 'even if a foreign accent is sometimes evident' well

> at A2 then it becomes 'noticeable foreign accent'. Well you could have a noticeable foreign accent that was sometimes evident so it would be better if it said 'is frequently evident' or 'is frequently heard' or something which then puts it, you're measuring it by how often do you hear it. You're measuring it in different ways because of that sort of language and that makes it difficult.

Thus, in contrast with the inherent ambiguity of a term like 'natural', the difficulty here comes in relation to distinguishing between the phrases 'noticeable' and 'sometimes evident', which may be synonymous in some understandings.

The comments regarding alternative descriptors in Excerpt 2.5 are interesting in light of the previous research reported above in which raters have critiqued the use of adverbs of frequency such as 'usually' because of their lack of concreteness. In Excerpt 2.5, we see that judging the same linguistic phenomenon (in this case, foreign accent) against statements expressed quite differently over levels is a challenging task. Against this critique, more consistent terminology with a quantitative slant is viewed as preferable. There are good methodological reasons to avoid describing a feature over levels by simply adjusting qualifiers to express degree (Council of Europe, 2001); however, in a scale of this kind which includes broad descriptions of pronunciation phenomena rather than the more discrete criterial features that might be identified through a data-driven scale design, the inclusion of some adverbs of frequency – such as 'rarely' or 'frequently' – may be a necessary compromise in order to separate otherwise synonymous terms.

Relevance

A third type of clarity problem related to relevance. This was not a significant topic in the focus group, but is worthy of discussion because it shows the consequences that perceived irrelevancies in a scale might have on rater behaviour. Excerpt 2.6 illustrates the difficulty raters had in interpreting the descriptor phrase 'pronunciation of a very limited repertoire of learnt words and phrases', which was viewed as out of step with the other scale descriptors.

Excerpt 2.6

Rater 3: But may I point out another problem for me with these descriptors is that in A1 there's a mention of very limited repertoire of learnt words and phrases which is grammatical. That doesn't seem to me to sit comfortably with the rest of the descriptors …

Several: Mm.

Rater 3: … and we already are trying to separate which as [Rater 4] says is quite difficult because you do judge, you don't just judge on what you hear you judge on the rest of the experience of the communication, and so you're struggling to actually separate pronunciation if it can be done from grammar and word choices and things … and yet *that* seems to fly in the face of all the rest of the descriptors.

There was evidence that this sort of confusion led raters to make their own decisions regarding the relevance of the descriptor, and this might have caused distortions in the rating process itself, with raters avoiding the descriptor altogether (as in Excerpt 2.7).

Excerpt 2.7

Rater 9: I mean because I did actually dismiss that first part of that descriptor because I didn't think it was relevant.

Rater 4: I didn't give any A1s because I felt that the students were speaking in sentences so that they didn't have a limited repertoire of learnt words and phrases. Also because they were speaking spontaneously and they could have gone any way with that you weren't just seeing what the pronunciation of a list of accepted words were, so I didn't go into that.

Conciseness

Despite the various problems with clarity identified above, there was one aspect of the scale which some raters perceived as positive: its brevity (see Excerpt 2.8).

Excerpt 2.8

Moderator: … were there any things that you liked about the scale in particular?

Rater 1: It only has six categories, which is good.

For Rater 1, brevity was a benefit when faced with the time constraints of judging real-time speech (although it should be noted that the task raters judged resulted in speech samples that were shorter than many oral proficiency style speaking exams).

Excerpt 2.9

Rater 1: Otherwise it's very difficult to decide in such a short time if you have too many [levels].

There was also an acknowledgement that, although the level of detail in descriptors was not sufficient in many cases, raters require a scale that is 'relatively simple' (see Excerpt 2.10).

Excerpt 2.10

Rater 4: But also I suppose if you're using this as a scale for assessment, I could see the problem if I was somebody trying to create a scale for other people to use that you can't have too many categories otherwise you're trying to think 'oh have I considered that?' … you've got to have something that's relatively simple, and um perhaps it's steered towards the side of being a bit too short.

This excerpt shows very effectively the tension that exists between the need for construct coverage and the requirement not to overburden the rater with complexity during the rating process. It is interesting to note that, although some raters criticized the scale for lacking detail in key areas (see Coherence), there were no requests for more than six levels, or for pronunciation to be assessed analytically rather than holistically.

Intuitiveness

'Intuitiveness' is here defined as the extent to which the scale might have anticipated the features of pronunciation that the raters attended to in rating the speech samples. The raters discussed this issue extensively, and this theme has been divided into discussions relating to (1) missing elements and (2) the capacity of the scale to capture the full construct that was perceived by raters.

Missing elements

The one missing element of the scale that raters discussed was the absence of 'self-repair' in the descriptors – an aspect that the raters found salient in their own perceptions of the speech samples (see Excerpt 2.11).

Excerpt 2.11

Rater 6: And also I noticed some self-repair …
Rater 3: Yes, yes.
Moderator: Oh, for mispronunciations?
Rater 6: Yes, so then where are we placing that? …
Rater 1: … if you're able to do that you should be already higher …
Rater 9: Well it depends on the frequency of it really … 'cause you can have endless self-correction …
Rater 4: It depends also on the level of the word you're correcting. If they're making a mistake with 'can' … or 'case' and they

> self-correct that, perhaps it's not the same as there was the example of somebody who said 'his accomplice' … okay now if they'd got that pronunciation wrong the first time and self-corrected you know I think we have a different idea of that don't we.

There are several useful lessons here which could inform the inclusion of self-repair in pronunciation descriptors (provided it was considered construct relevant). First is that self-repair is perceived by raters to be a higher level skill; secondly, it should be considered within reason – perhaps with reference to isolated incidents; thirdly, it might be lexically dependent, with candidates not penalized for attempting to pronounce words containing more difficult consonant clusters, for example. This example demonstrates clearly the sort of usable information that might be ascertained through consultation with experienced raters.

Capturing the perceived construct

Another key topic in the focus group was on the difficulty of assessing pronunciation in isolation from other elements of the speaking construct, particularly fluency but also grammar. Excerpt 2.12 presents a lengthy exchange between raters where elements that influenced pronunciation ratings beyond the construct embodied in the scale are discussed.

Excerpt 2.12

Rater 5: One of the things that I noticed and was rather struck by was that it wasn't necessarily the pronunciation of the individual phonemes that caused the largest, the interference with the L1 … it wasn't that that actually caused the difficulty in understanding, the difficulty in understanding which made me want to choose some of the lower ones was caused more by chunking – failure in chunking – and um also stress timing, syllable lengthening and the consonant cluster problems that the Chinese speakers have that cause them to make it choppy …

Several: Mm hm.

Rater 5: So although the pronunciation of individual phonemes does have an effect, it wasn't actually the thing that affected understanding …

Rater 1: That's right because I noticed exactly the same. You can understand separate words very well, they pronounce them quite well, but when they put them into a sentence they can't really say the whole sentence altogether …

Rater 7: Stress timing wasn't there.

Rater 5: And as you're listening your flow of understanding is interrupted by the choppiness or the pauses or the hesitation or the self-correction and so on. And those things seem to have a much more profound effect than individual phoneme pronunciation.
Rater 4: But then that is not always pronunciation, is it? I take your point, but that is also linked with …
Rater 5: Fluency.
Rater 4: The hesitation, grammar …
Several: Yes, grammar.
Rater 4: Sometimes I found it very difficult to take out the grammar because it was so closely linked, so I've left you some little notes whether you want them or not.

The difficulty in separating pronunciation out form other dimensions of the speaking construct was a key theme in Yates *et al.*'s (2011) study as well, and the further support for this issue in this study provides a strong argument that raters' views should be considered seriously. The theoretical divide between pronunciation and aspects of fluency – stress timing, hesitation, 'chunking' – becomes harder to maintain when human raters, who need to apply scales in practice, struggle to separate these dimensions for judgement purposes. Raters are conscious, too, of the role grammar appears to play in pronunciation judgements. Rater 4's reference to 'little notes' referred to comments made about individual speakers on the rating sheet such as 'grammar also affects understanding'. This came despite explicit instructions to raters that they should rate for pronunciation only, not for other dimensions of speech. It would seem, therefore, that the raters were more consciously oriented towards a broader comprehensibility construct which, as recent research indicates (e.g. Isaacs & Trofimovich, 2012; Saito *et al.*, 2015), is typically influenced by lexico-grammatical and fluency variables as well as segmental and prosodic features of L2 speech. The implications of this will be discussed in the Discussion section below.

Theoretical currency

The final theme, while brief and perhaps not representative of the group as a whole, is still noteworthy. One rater characterized the scale as 'outdated' in its view of English (see Excerpt 2.13):

Excerpt 2.13

Rater 5: It does seem to reflect a set of attitudes that come from some paradigms that are maybe becoming outdated, 'cause we're all having to adjust our own paradigms about what English is …

Moderator: Can you just elaborate on that a bit [Rater 5]?
Rater 5: I think we are being challenged to consider what is natural English these days. We're having to accept that there are Englishes that are accepted globally, and I think that this scale doesn't really represent the development of that thinking in the world of EFL.

This point connects with two other themes identified in the data. The first is the problem of coherence with the foreign accent descriptor. In talk around that issue, raters expressed feeling uncomfortable with the general idea of penalizing a foreign accent. One example from that discussion came from Rater 1 who was herself bilingual:

Excerpt 2.14

Rater 1: … it's much easier for me – for example – to understand an Italian speaker who *will speak like this* <does Italian accent> to understand than a native English speaker who might have a perfect accent.

Secondly, the term 'natural' in Excerpt 2.13 is again critiqued as reflecting a normative approach to English pronunciation. It was clear from these passages in the focus group that several raters did not feel comfortable with the view of English that was embodied in the scale, and that some raters at least saw it as out of step with more recent conceptualizations which have problematized the reliance on the native speaker model, and which have highlighted the pluricentricity of English (see Davies, this volume).

Discussion

This study set out to provide an exploration of the usability of a pronunciation rating scale: to gather data about raters' experiences which could feed back into thinking about the design of pronunciation rating scales. Specifically, the study set out to answer two research questions. (1) Which aspects of the CEFR Phonological control scale, if any, do raters problematize? (2) What implications can be drawn from raters' comments for the design of pronunciation scales generally?

Identified problems

The study identified, through the themes revealed in the qualitative analysis, a set of problems raters experienced in applying the CEFR phonological control scale to a set of sample performances. The problems were organized into themes and sub-themes, the content of which is summarized in Table 2.3.

Table 2.3 Summary of identified problems

Theme	*Specific problems*
(1) Clarity 1.1 *Scale coherence* 1.2 *Terminological precision* 1.3 *Relevance*	• Integral elements of pronunciation featured at some scale levels and not others (e.g. intonation) • Use of ambiguous terms (e.g. 'natural') • Terminology not distinct across levels (e.g. noticeable versus sometimes evident) • Inclusion of descriptor phrases relating to vocabulary knowledge
(2) Conciseness	• 'A bit' short and lacking in detail, but brevity appreciated
(3) Intuitiveness 3.1 *Missing elements* 3.2 *Capturing the perceived construct*	• Absence of self-repair in scale descriptors • Scale does not account for influence of fluency variables (e.g. chunking, hesitation) and grammar (i.e. scale does not adequately reflect the broader comprehensibility construct)
(4) Theoretical currency	• View of English (e.g. inclusion of foreign accent descriptor) is outdated

Principles

Using the identified problems as a basis, a number of inferences can be drawn for the future design of user-centred pronunciation rating scales, framed here in terms of technical recommendations and construct recommendations. However, they come with a caveat: the subject of this study has been a holistic pronunciation scale of a type common in the communicative assessment of pronunciation (see Harding, 2013). Therefore, these recommendations would generalize to scale design within this broad approach. It is assumed that other usability principles will apply to pronunciation scale design within other traditions. Also, these recommendations do not comprise a full set of guidelines for pronunciation scale construction. Rather, they should be understood as lessons learned from a usability study which might be applied to other scale design procedures.

Technical recommendations

(1) Include all assessed elements of pronunciation across rating scale levels (segmental and suprasegmental features). Avoid the assumption that suprasegmental information is only important at higher levels.
(2) Avoid abstract terminology such as 'natural', which requires reference to a scale-external standard and which may function as an implicit normative concept.

(3) Avoid incongruous references to other skill/knowledge areas (e.g. lexico-grammar) unless these are purposefully included across all levels (see construct recommendations).
(4) Maintain consistency of terminology across the scale to reduce the challenge for raters in following the trajectory of a feature across levels.
(5) Keep scales brief – six levels appear to be sufficient. Within level descriptors, one or two clauses per level may be considered underspecified, but many more than this would be problematic. Three to five clauses per level may be optimal (see also Council of Europe, 2001).

Construct recommendations

(1) Consider collapsing pronunciation and fluency into the same criterion. While there may be clear theoretical arguments to keep these sub-constructs separate within a broader speaking scale (see Brown & Fulcher, this volume), the raters here and in other studies (e.g. Yates *et al.*, 2011) have reported difficulty in delimiting a pronunciation construct which is not influenced by fluency factors such as stress-timing, hesitation and choppiness. Two useful examples of scales where pronunciation and fluency features are collapsed within the same criterion are the 'Delivery' dimension of the TOEFL speaking scale (Educational Testing Service, 2014) which includes reference to the intelligibility of articulation and intonation as well as to flow and pace, and the 'Delivery' criterion of the Trinity College London Integrated Skills in English examination, which covers intelligibility, lexical stress/intonation, fluency and effects on the listener in its descriptors (Trinity College London, 2015).
(2) While judgements of grammar also appear to be inseparable from some evaluations of pronunciation there will often be practical reasons for providing a separate scale criterion for spoken grammar in analytic speaking scales. In such cases, rater training should focus on techniques to deal with each element – grammar and pronunciation – in isolation, to the greatest extent possible. However, in stand-alone assessments of pronunciation where a comprehensibility construct is to be operationalized, the difficulty for raters in separating out grammatical and phonological features in forming judgements of ease of understanding will remain a key scale design issue, and grammatical accuracy will perhaps need to be encoded in descriptors across levels.
(3) Remove references to foreign accent in pronunciation scales unless there are clear purpose-driven reasons to assess strength of accent (rather than intelligibility). There are numerous studies that provide an evidence base to claim that the conflation of increasing intelligibility with decreasing foreign accent is untenable (see Munro, 2008). This study has also shown that the inclusion of foreign accent descriptors can be confusing for raters, and is perceived as anachronistic.

Limitations

The findings of this study are subject to two clear limitations. First, it must be acknowledged that there is an inherent circularity to the approach taken in selecting experienced raters to provide their views on usability. Many of the raters in the current study were experienced IELTS or Trinity College London examiners, and as such their views of the construct will have been influenced by their training and experience examining for these tests. Unless naïve raters (lay listeners, or teacher-raters not familiar with any particular rating scale) are used, this sort of circularity will be unavoidable. All experienced raters will have been apprenticed in a specific test and will have internalized these constructs. For the purposes of rating scale design projects, this will be less of a problem because the developer will need to ensure usability for a defined population, that is, a cohort of raters. For research purposes, however, it might be useful to carry out a study of this kind with naïve raters to remove the effect of prior scale experiences.

A second limitation relates to the use of the focus group, which has been critiqued in usability research within other fields (e.g. Nielsen, 1997). Focus groups may capture what users perceive as the strengths and weaknesses of a particular designed object or system, but they do not allow for direct observation of the individual actually using the object or system. To some extent, this criticism has been partially addressed in the current study by running the focus group directly after a rating session where use of the rating scale was still fresh in the raters' minds. Nevertheless, more fine-grained evidence would be very useful in future research in order to understand 'true' usability (see below).

Further research

This investigation into the usability of a pronunciation scale opens up interesting possibilities for further research. With a specific focus on pronunciation rating scale development, it would be useful to carry out what would be the logical next step to the current study: to integrate raters' suggestions into a revised instrument, re-run the rating session with the improved set of descriptors, and analyze the results. This would be a partial replication of the approach taken by Harsch and Martin (2012) for writing, which led to a scale that was strongly connected to the CEFR but which was also interpretable and usable for raters. It would also be of great benefit to the field to see more usability studies, which might be routinely conducted 'in-house', reported in the research literature. As this study has demonstrated, even scale-specific usability problems might have general relevance to the field.

There is also potential for a broader research programme exploring the usability of rating scales for all skill areas where language performance is judged. One option – following on from the critique made above concerning

the limitations of focus group research – would be to make use of eye-tracking methodology to fully understand how raters engage with specific scales: what they look at, when they look at it, what they don't look at, and how variations in scale complexity, layout and terminology might affect the ways in which raters interact with rating scales. Some research in this direction has already begun on the use of writing scales (see Winke & Lim, 2015). A second option would be to employ experimental methods where scales are modified to vary along dimensions of complexity, terminological concreteness, length, and so on. In this way, specific scale-related elements that affect rating behaviour, or that enhance usability, might be identified. Such features will no doubt vary according to individuals (e.g. Eckes, 2008), and so understanding the interaction between scale design and rater 'styles' will also be important. There is scope, in other words, for a more comprehensive type of scale usability research in language testing and especially in testing of L2 pronunciation.

References

Alderson, J.C. (1991) Bands and scores. In J.C. Alderson and B. North (eds) *Language Testing in the 1990s* (pp. 71–86). London: Modern English Publications/British Council/Macmillan.

Baker, B.A. (2012) Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly* 9 (3), 225–248.

Bejar, I. (2012) Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice* 31 (3), 2–9.

Braun, V. and Clarke, V. (2006) Using thematic analysis in psychology. *Qualitative Research in Psychology* 3 (2), 77–101.

Brindley, G. (1998) Describing language development? Rating scales and SLA. In L.F. Bachman and A.D. Cohen (eds) *Interfaces between Second Language Acquisition and Language Testing Research* (pp. 112–140). Cambridge: Cambridge University Press.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council of Europe (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. A Manual*. Strasbourg: Language Policy Division.

Cumming, A., Kantor, R. and Powers, D.E. (2002) Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal* 86 (1), 67–96.

Eckes, T. (2008) Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing* 25 (2), 155–185.

Educational Testing Service (2014) Independent speaking rubrics/integrated speaking rubrics. See http://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf.

Fulcher, G. (1996) Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13 (2), 208–238.

Fulcher, G. (2003) *Testing Second Language Speaking*. London: Pearson Longman.

Galaczi, E.D., Ffrench, A., Hubbard, C. and Green, A. (2011a) Developing assessment scales for large-scale speaking tests: A multiple-method approach. *Assessment in Education: Principles, Policy & Practice* 18 (3), 217–237.

Galaczi, E., Post, B., Li, A. and Graham, C. (2011b) Measuring L2 English phonological proficiency: Implications for language assessment. In J. Angouri, M. Daller and

J. Treffers-Daller (eds) *The Impact of Applied Linguistics: Proceedings of the 44th Annual Meeting of the British Association for Applied Linguistics* (pp. 67–72). London: Scitsiugnil Press.

Harding, L. (2013) Pronunciation assessment. In C.A. Chapelle (ed.) *The Encyclopedia of Applied Linguistics.* Oxford: Blackwell.

Harding, L. (forthcoming) Comprehensibility or nativeness? Investigating how teacher-raters interpret and apply a pronunciation rating scale.

Harding, L., Pill, J. and Ryan, K. (2011) Assessor decision making while marking a note-taking listening test: The case of the OET. *Language Assessment Quarterly* 8 (2), 108–126.

Harsch, C. and Martin, G. (2012) Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing* 17 (4), 228–250.

Heaton, J. (1975) *Beginning Composition through Pictures*. Harlow: Longman.

Horner, D. (2013) Towards a new phonological control grid. In E. Galaczi and C. Weir (eds) *Exploring Language Frameworks: Proceedings of the ALTE Kraków Conference, July 2011*. Cambridge: Cambridge University Press.

IELTS (2015) *IELTS Guide for Teachers: Test Format, Scoring and Preparing Students for the Test*. See https://www.ielts.org/~/media/publications/guide-for-teachers/ielts-guide-for-teachers-2015-uk.ashx (accessed 23 July 2016).

Isaacs, T. (2013) Assessing pronunciation. In A.J. Kunnan (ed.) *The Companion to Language Assessment* (pp. 140–155). Hoboken, NJ: Wiley.

Isaacs, T. and Trofimovich, P. (2012) Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition* 34 (3), 475–505.

Isaacs, T., Laurier, M.D., Turner, C.E. and Segalowitz, N. (2011) Identifying second language speech tasks and ability levels for successful nurse oral interaction with patients in a linguistic minority setting: An instrument development project. *Health Communication* 26 (6), 560–570.

Isaacs, T., Trofimovich, P., Yu, G. and Chereau, B.M. (2015) Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS Research Reports Online* 4, 1–48.

Knoch, U. (2009) *Diagnostic Writing Assessment: The Development and Validation of a Rating Scale*. Frankfurt am Main: Peter Lang.

Levis, J.M. (2005) Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly* 39 (3), 369–377.

May, L. (2009) Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing* 26 (3), 397–421.

Mcnamara, T. (1996) *Measuring Second Language Performance*. London: Longman.

Munro, M.J. (2008) Foreign accent and speech intelligibility. In J.G. Hansen Edwards and M. Zampini (eds) *Phonology and Second Language Acquisition* (pp. 193–218). Amsterdam: John Benjamins.

Nielsen, J. (1997) The use and misuse of focus groups. *IEEE Software* 14 (1), 94–95.

Norman, D.A. (2002) *The Design of Everyday Things*. New York: Basic Books.

North, B. (2000) *The Development of a Common Framework Scale of Language Proficiency*. New York: Peter Lang.

North, B. (2014) *The CEFR in Practice*. Cambridge: Cambridge University Press.

Orr, M. (2002) The FCE speaking test: Using rater reports to help interpret test scores. *System* 30 (2), 143–154.

Pearson Education (2012) PTE General score guide. See http://pearsonpte.com/wp-content/uploads/2014/07/PTEG_ScoreGuide.pdf.

Ryan, K. (2007) Assessing the OET: The nurses' perspective. Unpublished manuscript, University of Melbourne.

Saito, K., Trofimovich, P. and Isaacs, T. (2016) Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics* 37, 217–240.

Stake, R. (1995) *The Art of Case Study Research*. Thousand Oaks, CA: Sage.

Trinity College London (2015) Integrated Skills in English (ISE) specifications: Speaking & listening. ISE Foundation to ISE III. See http://www.trinitycollege.com/resource/?id=6298.

Turner, C.E. and Upshur, J.A. (2002) Rating scales derived from student samples: Effects of the scale marker and the student sample on scale content and student scores. *TESOL Quarterly* 36 (1), 49–70.

Upshur, J.A. and Turner, C.E. (1995) Constructing rating scales for second language tests. *ELT Journal* 49 (1), 3–12.

Van Moere, A. (2013) Raters and ratings. In A.J. Kunnan (ed.) *The Companion to Language Assessment* (pp. 1358–1374). Hoboken, NJ: Wiley.

Winke, P. and Lim, H. (2015) ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing* 25, 37–53.

Yates, L., Zielinski, B. and Pryor, E. (2011) The assessment of pronunciation and the new IELTS Pronunciation Scale. In J. Osborne (ed.) *IELTS Research Reports* 12 (pp. 1–46). Melbourne and Manchester: IDP IELTS Australia and British Council.

Part 2

Insights From Assessing Other Language Skills and Components

3 Pronunciation and Intelligibility in Assessing Spoken Fluency

Kevin Browne and Glenn Fulcher

Introduction

This chapter argues that any definition of the construct of fluency must include familiarity of the listener with the entire context of an utterance. This extends to pronunciation, the intelligibility of which is an interaction between the phonological content of the utterance and the familiarity of the listener with the second language (L2) speech produced by speakers from a specific first language (L1) background. This position recognizes that successful communication is not merely a matter of efficient cognitive processing on the part of the speaker. Fluency is as much about perception as it is about performance. This is a strong theoretical stance, which can be situated within an interactionist perspective on language use. Good theory generates specific predictions that may be empirically tested. If the listener is critical to the construct, we would expect to discover two facts. First, that variation in listener familiarity with L2 speech results in changes to scores on speaking tests. Secondly, that this variation is associated with estimates of intelligibility when the speaker is kept constant. In this chapter we describe a study that investigates these two predictions. We situate the findings in the context of language testing, where variation in familiarity among raters is a cause for concern.

The Fluency Construct

The construct of fluency is endemic in language teaching and applied linguistics research. Teachers feel especially relaxed in using the term to refer to a desirable quality of learner speech that approximates 'nativelike delivery' – or

'proficiency' in the broadest sense (Lennon, 1990). This comfortable assumption hides the fact that there is no single definition of 'nativelike' within a single language (Davies, 2004), and variation between languages is frequently considerable (Riazantseva, 2001). Early research by Fillmore (1979) and Brumfit (1984) provided a very broad definition of fluency, including 'filling time with talk' through automatized language production, selecting relevant content for context, and creating coherent utterances without becoming 'tongue tied'. Koponen and Riggenbach (2000) exposed the metaphorical nature of the fluency construct, characterizing speech as fluid, or flowing like a river: smooth and effortless in its passage from mind to articulation. The articulation includes pronunciation, which adds or subtracts from the perception of fluidity (Educational Testing Service, 2009) on the part of the listener.

The language of fluency definitions reveals what we have elsewhere called the 'janus-faced' nature of the construct (Fulcher, 2015: 60). Language testers often make the assumption that pronunciation is a simple 'on/off switch' for intelligibility (Fulcher, 2003: 25). But this assumption focuses too much upon the production of the individual speaker in relation to the acquisition of some standard, usually the notion of the 'native speaker'. It is the assumption that underlies the automated assessment of pronunciation in computer-based tests by matching performances on reductive task types such as sentence repetition (Van Moere, 2012) with preselected norms. The place of pronunciation in cognitive fluency models also treats phonological accuracy as merely the observational component of part of a speech-processing model such as that of Levelt (1989, 1999), so that measurements may be treated as surrogates for general L2 proficiency (Segalowitz, 2010: 76).

The reality is that pronunciation is variably problematic, depending on the familiarity of the listener with the L1 of the speaker. This realization is significant in the context of language assessment, where such familiarity becomes an important variable that impacts scores being assigned to speakers.

Defining intelligibility and familiarity

Familiarity shapes and facilitates speech processing. The intelligibility of speech is speaker–listener dependent (Riney *et al.*, 2005). Attention has been drawn to how differential rater familiarity with accent can affect test scores, posing a threat to both reliability and validity (e.g. Carey *et al.*, 2011; Winke *et al.*, 2013; Xi & Mollaun, 2009). Research into rater accent familiarity as a potential threat has tended to focus on listeners' shared L1 with the test takers (Kim, 2009; Xi & Mollaun, 2009), residency and employment in the country where the L1 of test takers is spoken (Carey *et al.*, 2011), and prior personal L2 study experiences (Winke *et al.*, 2013). In these studies the

construct of familiarity was not carefully defined, but was inferred on the basis of different types and amounts of linguistic experiences a rater had with the L2 accent. A definition that can be extrapolated from these studies is that accent familiarity is a speech perception benefit developed through exposure and linguistic experience. Carey *et al.* (2011: 204) labelled it 'interlanguage phonology familiarity'.

Gass and Varonis (1984) released the earliest study of familiarity. They argued that four types of familiarity contribute to comprehension: familiarity with topic of discourse; familiarity with nonnative speech in general; familiarity with a particular nonnative accent; and familiarity with a particular nonnative speaker. Their study used 142 native-speaking university students as participants who listened to recordings of two male Japanese-English speakers and two male Arabic-English speakers completing three reading-aloud tasks: (1) reading a story; (2) reading a set of five 'related sentences' that pertained to the story although not included in the text; and (3) a set of 'unrelated sentences' with contexts or topics pertaining to 'real world knowledge'. The recordings were used to create 24 different 'tapes'. Each tape included first a reading of either the 'related' or 'unrelated' sentences. Next came a reading of the story, followed by the set of sentences not included prior to the story. The items were read by different combinations of speakers. Each listener was asked to complete transcription tasks of the related and unrelated sentences, and produce a short summary of the story as a measure of comprehension.

Gass and Varonis concluded that 'familiarity of topic' is the greatest contributor to comprehension of the four familiarity types researched (see also Kennedy & Trofimovich, 2008). This was determined by one-tailed *t*-tests comparing the pre- and post-text transcriptions of the related sentences. The results revealed a significant difference of means of errors ($p < 0.05$) for three of the four speakers (Gass & Varonis, 1984: 72). More errors were reported in the pre-story transcriptions of the 'related' sentences than in the post-story transcriptions, suggesting that native speakers are more capable of determining the content of nonnative speakers' utterances if they know the specific topic. Likewise, the 'unrelated' sentences determined to be comprised of 'real world knowledge' resulted in a significantly lower instance of errors ($p < 0.0001$) as compared to the 'related' sentences when they occurred in the pre-story position on the tapes.

Familiarity of speaker, familiarity of accent and familiarity of nonnative speech in general were found to contribute to the comprehensibility of nonnative speakers, although these findings were not based on any statistically significant differences in the data. Familiarity of accent was determined to positively affect transcription accuracy by observing instances of speaker error in the pre- and post-story positions. Greater accuracy was observed when listeners had encountered the same accent in the pre-story or story reading when transcribing the post-story sentences.

It can be argued that what Gass and Varonis discovered was that familiarity facilitates 'intelligibility' and not 'comprehension', according to the more useful definitions provided by Smith and Nelson (1985: 334). Smith and Nelson suggested the following interpretations of intelligibility, comprehension and interpretability:

- *intelligibility*: word/utterance recognition;
- *comprehensibility*: word/utterance meaning (locutionary force);
- *interpretability*: meaning behind word/utterance (illocutionary force).

Although Gass and Varonis did include the story summary for listener participants there was no analysis or discussion of the data to support the claim that the different types of familiarity they examined contribute to comprehension, which would include out of necessity the notions of locutionary or illocutionary force. While we do not wish to argue against the possibility that familiarity may contribute to comprehension and determination of meaning, Gass and Varonis' findings can only be said to relate to intelligibility of word or utterance recognition, depending upon listener familiarity.

As Smith and Nelson (1985: 334) suggested, the terms 'intelligibility', 'comprehension' and 'interpretability' should be defined to avoid confusion, since these terms have been applied in various ways and at times interchangeably. The definition of intelligibility in this research follows Field (2005), as being how the phonological content of a speaker is recognized by the listener. This definition takes into account how the listener processes utterances, which we argue is a function of level of familiarity.

It is therefore theorized that increasing accent familiarity reduces the processing effort required for the phonological content of speech. Thus, raters with higher levels of familiarity are more likely to find speech intelligible, while lower levels of familiarity reduce intelligibility. Familiarity on the part of the listener is therefore the most important variable to impact the intelligibility aspect of fluency, which results directly in score variation (Derwing *et al.*, 2004).

Research questions

In order to investigate the role of intelligibility as a critical component of fluency within the argument that the construct exists as much within the listener as it does within the speaker, we formulated two research questions:

(1) How do raters' familiarity levels with L2 English spoken by L1 speakers of Japanese affect pronunciation test scores?
(2) How do raters' familiarity levels with L2 English spoken by L1 speakers of Japanese affect intelligibility success rates?

Methodology

No previous study of rater accent familiarity as a threat to test validity has simultaneously examined how raters score candidates on operational tests concurrently with rater intelligibility success rates. As a result, little is known about why score differences occur. The methodology therefore provides the means to investigate the relationship between these variables and identify potential impact on scores.

Participants

Eighty-seven ESL/EFL teachers and/or graduate students enrolled in applied linguistics or TESOL programmes were recruited via email to participate as volunteer rater participants. Most ($n = 73$) were L1 English speakers and 14 were L2 speakers (see Table 3.1).

Five first-year Japanese university students studying English as non-English majors at Tsukuba University (male: $n = 1$; female: $n = 2$), Waseda University (male: $n = 2$), and one American male from the Southern United States were recruited as the speaker participants. The students were enrolled in intermediate-level English courses at the time, and had studied English for six years prior to participating.

Table 3.1 Rater participants' home country list

United Kingdom	35
USA	34
Canada	7
South Africa	4
Japan	4
Australia	3
Brazil, France, Jamaica, Libya, Malta, Spain, St. Lucia, Sudan, Syria, Ukraine	1 (per country)
Total	87

The test

A three-part test was constructed to measure rater intelligibility success rates for comparison with the scores allocated to different speakers. Since participation was voluntary, the test was designed to be completed in less than 25 minutes. Rater participants required a computer connected to the internet and were recommended to complete the test with headphones in a quiet room.

Part 1 of the test included questions related to raters' professional, biographical and linguistic experiences. Questions focused on their L1(s), home

country, country of residence at that time, ESL/EFL teaching and/or research experience, and familiarity with Japanese-English. Raters' familiarity with the accent was determined from responses to a four-level self-reporting scale. The scale and number of participants selecting each level was:

- **No familiarity** (*n* = 13).
- **Limited familiarity**: You have heard Japanese speakers of English but without regularity, and/or have not had Japanese students during the last two years (*n* = 32).
- **Some familiarity**: You have spent at least the last two years with students from Japan, have visited Japan and/or regularly watch TV or movies in Japanese (*n* = 4).
- **Very familiar**: You are a native speaker of Japanese, have lived in Japan for one or more years and/or have studied the Japanese language for one or more years (n = 38).

Part 2 was divided into six sections, with one section for each speaker participant. Each section contained a recording of the speaker reading two sentences. The raters were asked to listen to each sentence and then complete an intelligibility gap-fill task by typing missing words from an incomplete transcript of the sentences on the screen. The native speaker was placed in first position. This was decided primarily to help the raters better understand the tasks they were asked to complete, and to serve as an 'easily intelligible' example of pronunciation to process. There were a total of 28 intelligibility gap-fill items in the test (24 spoken by the Japanese-English speakers; four spoken by the native speaker).

After completing the intelligibility task for one speaker, raters scored that speaker for pronunciation using a five-point scale adapted from the TOEFL iBT Speaking Scoring Rubric 'Delivery' sub-scale for the independent speaking tasks, which incorporates the notion of 'fluidity' (Educational Testing Service, 2009). The scale that the raters used in the current study is shown in Table 3.2. Each recording was approximately 18 seconds in length. Raters could start, stop or replay the recording at their discretion. No visuals were provided; raters had no additional information about the speakers that would lead to inferences that might impact scores (e.g. gender, age, L1, nationality) (see Rubin, 1992). There are a number of limitations in the methodology. First, raters completed test items in the same sequence. The survey website made randomizing the items prohibitive, as they were clustered according to speaker, so order effect could not be controlled. Secondly, the native speaker may have 'loomed over the study' (Isaacs & Thomson, 2013), but none of the raters reported the use of a native speaker example to have been problematic, and the data from the native speaker were not included in the analyses.

The sentences read by the speaker participants were adaptations of the Bamford–Kowal–Bench (BKB) sentence lists (Bench *et al.*, 1979), which were

Table 3.2 Pronunciation score descriptors used in the current study

5	Speech is generally clear and requires little or no listener effort. Only one listening required.
4	Speech is generally clear with some fluidity of expression, but it exhibits minor difficulties with pronunciation and may require some listener effort at times. Only one listening required.
3	Speech is clear at times, although it exhibits problems with pronunciation and so may require more listener effort. It was necessary to listen more than once before attempting to complete the gap fill.
2	Consistent pronunciation difficulties cause considerable listener effort throughout the sample. It was necessary to listen more than once before attempting to complete the gap fill.
1	Cannot comprehend at all.

Source: Adapted from the TOEFL iBT Speaking Scoring Rubric, Independent Tasks (Educational Testing Service, 2015: 189–190).

originally designed to measure the listening capabilities of children with varying degrees of sensorineural hearing loss. Sensorineural hearing loss is an affliction that affects how speech is processed. Regardless of the volume of the speech signal, sensorineural hearing loss affects the clarity of the acoustic signal the listener perceives. Like Bench *et al.*'s original tests, this test was designed to measure differences in speech perception and processing with gap-fill transcription tasks with clarity of speech determined through word identification accuracy.

The BKB test measures speech perception abilities using samples with pronunciation a 'normal' listener should find intelligible, whereas the test designed for the research described in this chapter measures speech perception using accented samples for which the rater participants had variable familiarity. The BKB sentences were standardized in length and lexical complexity and served to reflect natural speech of NS children aged 8–15 (see Table 3.3). The sentences designed for this study were also standardized in length and lexical complexity to represent the vocabularies of intermediate-level Japanese-English speakers. Lexical complexity was determined utilizing the JACET

Table 3.3 Examples of the original BKB sentences

An old woman was at home.
He dropped his money.
They broke all the eggs.
The kitchen window was clean.
The girl plays with the baby.

Source: Bench *et al.* (1979: 109)

8000, a corpus of the 8000 most frequently used English words by Japanese speakers of English. Lexical complexity was restricted to the 3000 most frequently used words in order to eliminate the need to provide explanations of word meaning or pronunciation to speaker participants. As a result, each speaker was left to pronounce each word in a sentence as they thought fit.

A unique aspect of the sentences designed for this instrument was the decision to intentionally construct them to have complex or unpredictable contexts. As previously discussed, Gass and Varonis (1984) argued that 'familiarity of context' was the most significant contributory type of familiarity to success in word/utterance identification tasks. This is because background knowledge of context helps the listener to successfully guess words or utterances that he or she is not able to otherwise identify. We judged that the use of sentences with complex or unpredictable contexts might effectively reduce the context familiarity benefit identified by Gass and Varonis, thus allowing us to see the impact of pronunciation alone on listener evaluation of intelligibility. The resulting sentences constructed for the test were not nonsensical; they were syntactically accurate although contextually complex or unpredictable (see Table 3.4).

The sentences were also designed to feature aspects of Japanese-English phonology that are known to be problematic both in production for the speakers and in distinction by unfamiliar listeners. Elements of problematic Japanese-English phonology incorporated in the test included /r/–/l/ distinction, the lax vowels /I/, /ʊ/, /ʌ/ and /ə/, and the voiced dental fricative /ð/ (see Carruthers, 2006, for a complete discussion of pronunciation difficulties of Japanese speakers of English).

Part 3 of the test sought rater comments in order to gain additional insight into the raters' opinions of the research instrument and their experiences completing the test.

Table 3.4 The test sentences developed for the current study

Speaker 1	They had a tiny day.
	The old soaps are dirty.
Speaker 2	They are paying some bread.
	The play had nine rooms.
Speaker 3	The institution organism was wet.
	The dog made an angry reader.
Speaker 4	The ladder is across the door.
	He cut his skill.
Speaker 5	The union cut some onions.
	She sensed with her knife.
Speaker 6	Mine took the money.
	The matches lie on the infant.

Analyses

Facets 3.71 Many Faceted Rasch Measurement (MFRM) software (Linacre, 2013) and SPSS (Version 20) were used to analyze the test data. MFRM allows for multiple aspects facets of a test to be examined together and, in the case of this study, to investigate raters' intelligibility scores and their abilities to transcribe utterances. Only data from the five L1 Japanese speakers were included in the MFRM analyses. This was designed to determine whether rater accent familiarity differences resulted in significant score differences. The pronunciation score and intelligibility success rates data were analyzed separately (as recommended by Linacre, personal communication) due to the differences of tasks, as fit statistics were compromised when the different tasks were analyzed together.

Two facets (the raters and speakers) and one grouping facet (raters' familiarity level with Japanese-English) were examined. The intelligibility data were also analysed examining two facets – the raters and the items – again with familiarity level as a grouping facet.

Findings and Discussion

MFRM analyses of the pronunciation scores yielded results supporting previous findings that raters' familiarity with speakers' accents can have a significant effect on oral proficiency scores (e.g. Carey *et al.,* 2011; Winke *et al.*, 2011, 2013). The most informative and important piece of output from Facets analyses is the variable map, which summarizes the key information of each facet and grouping facet into one figure. The scale utilizes measurements in terms of 'logits' that reflect probability estimates on an equal-interval scale. Figure 3.1, which presents the Facets variable map for pronunciation scores, is separated into five vertical columns:

(1) Column 1 displays the logit scale ranging from –7 to 2. The scale provides a reference for measurements of all other columns. The measure 0 represents even likelihood, or 50–50 odds of prediction.
(2) The second column displays the leniency of each rater from most (top) to least.
(3) The third column shows the grouping facet revealing that the 'very familiar' group of raters were most lenient in scoring pronunciation.
(4) Column 4 shows the ability measures of each speaker-participant. The most proficient was speaker E shown at the top.
(5) The fifth column shows the five-point rating scale used to score pronunciation. Each speaker participant's position in the fourth column is horizontal to their mean score on this rating scale.

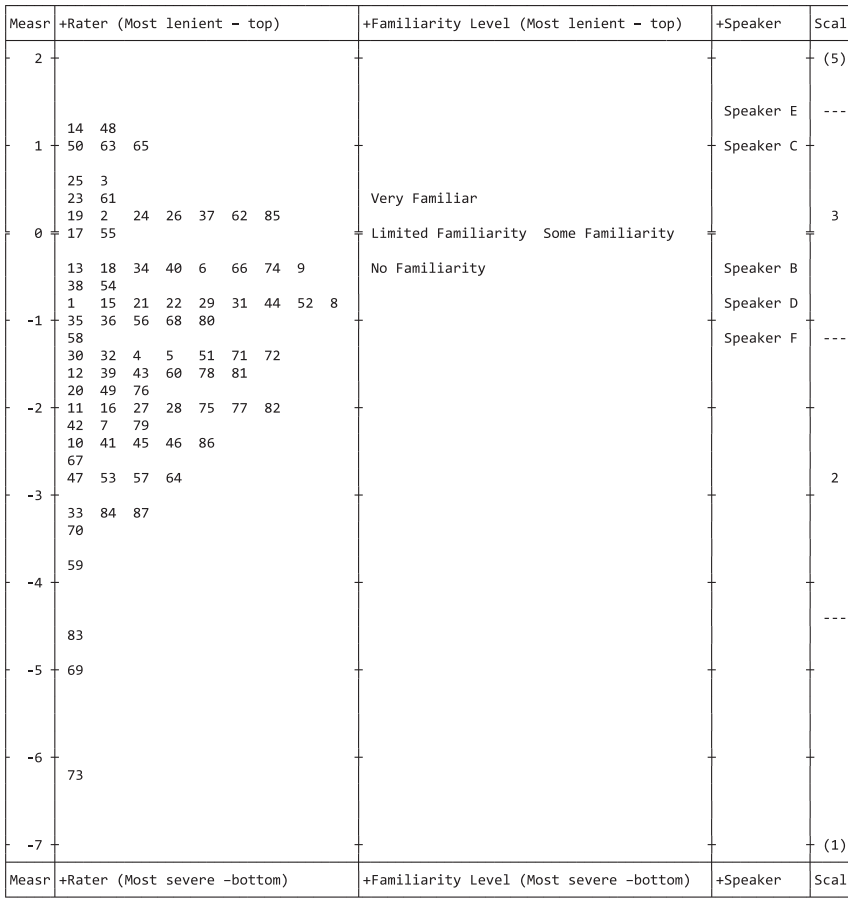


Figure 3.1 Facets variable map of pronunciation scores including four levels of familiarity

The range of rater severity shown in Column 2 (7.39 logits) is wider than the spread of speaker-participants' ability in Column 4 (2.7). This indicates that the individual differences of rater severity were high. A closer examination of rater performance by familiarity level provided in Table 3.5 indicates that, as familiarity level increases, so do scores and rater leniency. Pearson's chi-square indicates significant differences of pronunciation scores between the four groups ($\chi^2(3) = 12.3$, $p = 0.01$).

Figure 3.2 shows the correlation between level of familiarity and pronunciation score. Although shared variance is a modest 15%, in a speaking test such an influence may make a significant impact on an individual score.

The Facets variable map for intelligibility scores is shown in Figure 3.3. The content of each column is as follows:

Table 3.5 Pronunciation score: Facets rater familiarity level group measures

Familiarity level	*Total score*	*Obs. Av.*	*Measure in logits*	*Model SE*	*Infit MnSq*	*ZStd*
No	144	2.22	−0.38	0.22	0.91	−0.4
Some	48	2.4	−0.09	0.38	0.71	−0.9
Limited	388	2.47	0.03	0.14	0.92	−0.6
Very	526	2.77	0.43	0.12	1.09	0.9
Mean	276.5	2.46	0	0.21	0.91	−0.3
SD	190	0.2	0.29	0.1	0.14	0.7

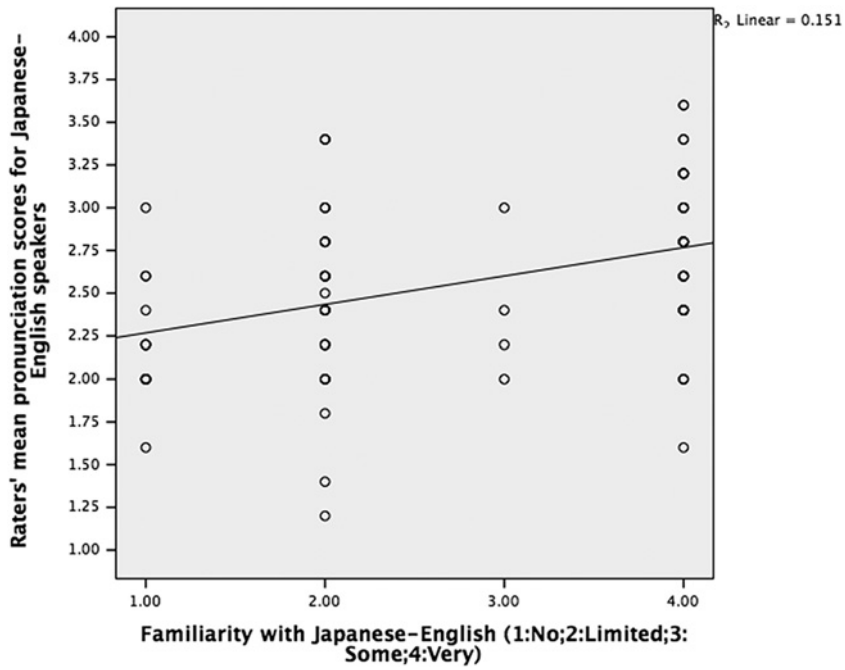


Figure 3.2 Scatterplot showing the correlation between accent-familiarity level and pronunciation scores

(1) Column 1 displays the logit scale ranging from −4 to 6.
(2) The second column shows how the individual raters performed in the intelligibility gap-fill exercises. Raters' individual abilities are reflected in their position on the map with the highest scoring raters at the top.
(3) The third column reveals how rater groups performed. As predicted, the 'very familiar' raters were the most successful and the 'no familiarity' group the least successful at completing the tasks.

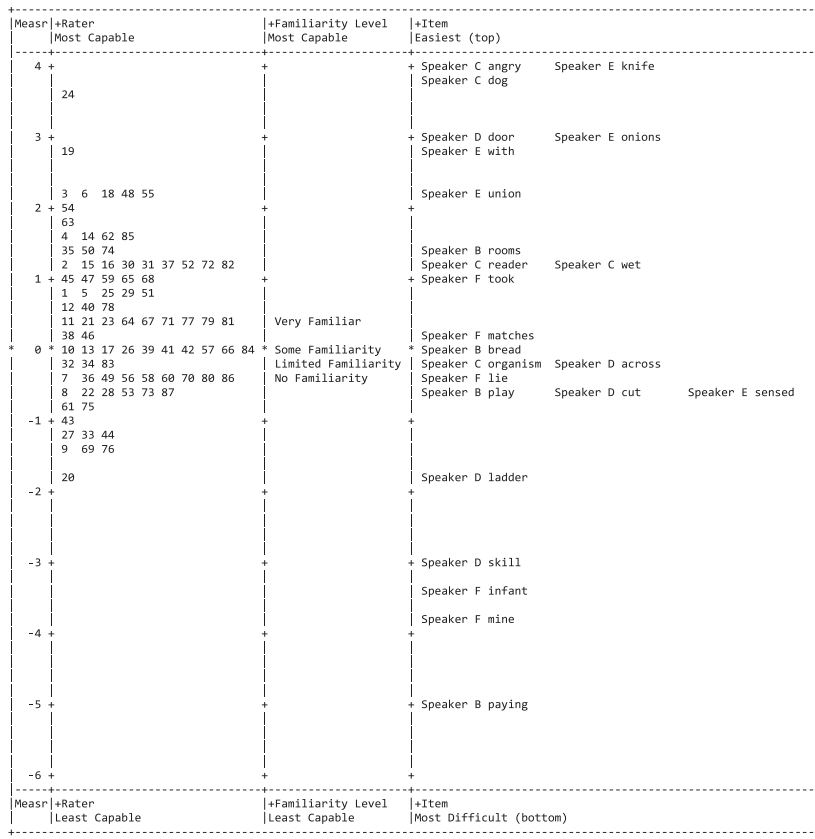


Figure 3.3 Facets variable map of intelligibility gap-fill outcomes including four levels of familiarity

(4) The fourth column displays the items from easiest (top) to most difficult (bottom). The items are identified first according to the speaker from whose recording they originated, and the target word. The column reveals that all five speakers produced items that were both easier (with logit scores above zero) and more difficult (with negative logit scores).

The most important results in Column 3 show that the more familiar raters are with Japanese English the more capable they are at transcribing the speakers' utterances. Table 3.6 shows that as familiarity with Japanese-English increases, so does observed intelligibility. Raters 'very familiar' with Japanese-English were 20% more successful than the raters with 'no familiarity'. Figure 3.4 shows that the correlation of the two variables share 31% variance, which indicates a potentially large impact of familiarity on intelligibility.

Table 3.6 Facets intelligibility familiarity level measurements

Familiarity level	*Total score*	*Total count*	*Obs. Av.*	*Measure in logits*	*Model SE*	*Infit MnSq*	*ZStd*
No	156	312	0.50	−0.41	0.16	0.99	0.0
Limited	435	768	0.57	−0.13	0.10	0.91	−1.7
Some	59	96	0.61	0.07	0.30	0.84	−1.0
Very	634	912	0.70	0.46	0.10	1.08	1.4
Mean	321.0	522	0.59	0.00	0.17	0.96	−0.4
SD	227.4	331	0.07	0.32	0.08	0.09	1.2

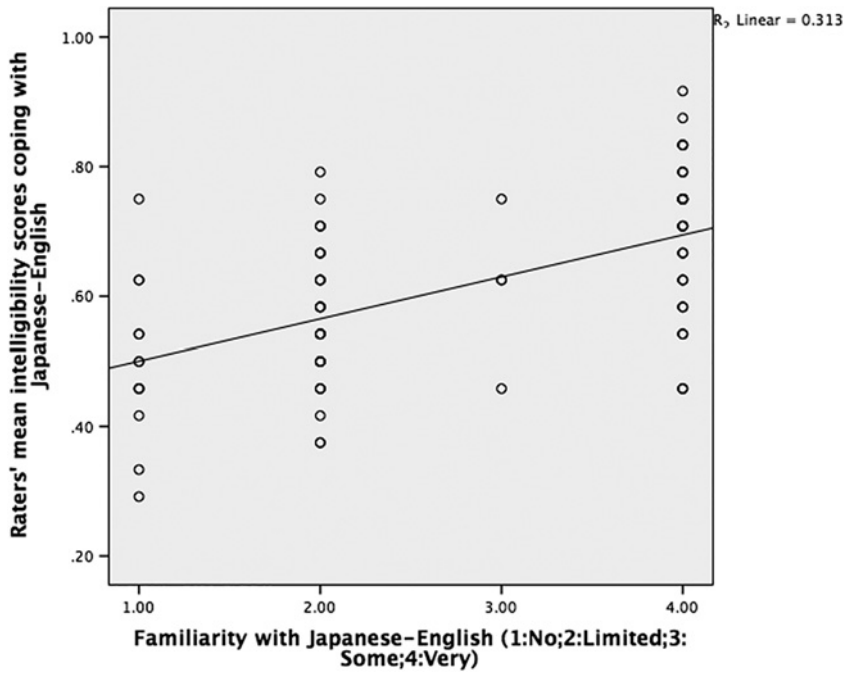


Figure 3.4 Scatterplot showing the correlation between accent-familiarity and intelligibility

Conclusion

We have argued that an understanding of fluency, and the place of pronunciation within a model of fluency, must take into account the listener. The study reported in this chapter addresses the two empirical correlates of the theoretical stance taken. The findings show that both pronunciation test

scores and intelligibility vary as a function of listener familiarity. While the current study focuses on pronunciation as one component of fluency, the study supports the theoretical stance that the construct of fluency more generally, and intelligibility more specifically, is situated as much within the listener as the speaker. Perhaps the reason for the listener being ignored in recent cognitive research is the absence of the listener from models of cognitive processing, such as that of Levelt, where it is argued that there are two major parts to speech processing:

> … a semantic system which 'map[s] the conceptualization one intends to express onto some linear, relational pattern of lexical items' and a phonological system which 'prepare[s] a pattern of articulatory gestures whose execution can be recognized by an interlocutor as the expression of … the underlying conceptualization'. (Levelt, 1999: 86)

A speech-processing model of this kind is typically represented as a flowchart. It therefore represents a 'software-solution' to the problem of mind and language. Taken literally, the interlocutor is relegated to the role of a passive recipient of the speaker's output, for which the speaker is completely responsible.

This is a convenient place to be if one wishes to use automated speech assessment systems, as the construct does not involve a listener, and the use of monologic and semi-direct tasks is rendered unproblematic. It could also be argued that listener variability is little more than error, which is eliminated by the removal of variable human raters in automated assessment (Bernstein *et al.*, 2010). However, if listeners are part of the construct, it would seem unreasonable to eliminate them from the equation completely. Language, after all, is a tool for human communication, and so it makes a difference who you are talking to, the context in which you are talking, and the purpose of the communication.

What this research does not do is identify a 'familiarity threshold' that might be recommended for a particular type of speaking test. What it does do is to argue that familiarity is inevitably part of the construct, and to problematize the relationship between familiarity, intelligibility and test scores for the purposes of assessing speaking. This is likely to be of particular importance in contexts where single raters are asked to rate the L2 speech of test takers drawn from a large variety of L1 backgrounds. This situation is common in large-scale L2 testing, where at present there is no attempt to match raters with speakers on the basis of rater familiarity with accented L2 pronunciation from the L1. The issue for high-stakes speaking assessment is the principle that construct-irrelevant facets of a test should be a matter of indifference to the test taker. The principle implies that the test taker should get a similar score (given random error) no matter which rater is randomly selected from the universe of raters available for selection. We normally refer

to this as the generalizability of the score across facets of the test (see Schoonen, 2012).

The discovery that the construct resides in the listener as much as in the speaker therefore leads to a dilemma: should familiarity be controlled in order to retain generalizability and the principle of equal treatment, or should familiarity be allowed to vary (as at present) as it is construct relevant? The problem is that although we have argued that familiarity is construct relevant, scores vary with familiarity. Unless it is possible to specify the level of familiarity that would be expected in the target domain to which test scores are intended to predict performance, it would seem reasonable to expect at least a minimum level of familiarity. This is certainly the case in large-scale tests that are used for a variety of decision-making purposes. Achieving familiarity may be obtained in one of two ways: first, by using a measure of familiarity such as the one used in this study to match raters with test takers; and secondly, by providing accent familiarity training to raters across the range of L1s represented in the test taker population at large. Further research is also required into the levels of rater familiarity required for there to be no impact on scores from intelligibility. Such research may need to have wider scales of familiarity than that used in this research, and have a much larger *n*-size for each L1 population, in order to maximize reliability. A larger study may be able to identify a plateau on the scale, which could then be used in conjunction with rater training to select raters for use with test takers from specific L1 backgrounds.

The salience of test method facets in score variance has always been one of the main considerations in investigating the fairness of decision making. It becomes even more problematic when the variance is construct relevant, but potentially random depending on how raters are selected. This paper problematizes the issue of potentially unfair construct-relevant variance, and points the way forward to potential remedies and future research.

References

Bench, J., Kowal, A. and Bamford, J. (1979) The BKB (Bamford–Kowal–Bench) sentence lists for partially hearing children. *British Journal of Audiology* 13, 108–112.

Bernstein, J., Van Moere, A. and Cheng, J. (2010) Validating automated speaking tests. *Language Testing* 27 (3), 355–377.

Brumfit, C. (1984) *Communicative Methodology in Language Teaching: The Roles of Fluency and Accuracy*. Cambridge: Cambridge University Press.

Carey, M.D., Mannell, R.H. and Dunn, P.K. (2011) Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing* 28 (2), 201–219.

Carruthers, S.W. (2006) Pronunciation difficulties of Japanese speakers of English: Predictions based on a contrastive analysis. *HPU TESL Working Paper Series* 4, 17–23.

Davies, A. (2004) The native speaker in applied linguistics. In A. Davies and C. Elder (eds) *The Handbook of Applied Linguistics* (pp. 431–450). London: Blackwell.

Derwing, T., Rossiter, M., Munro, M. and Thomson, R. (2004) Second language fluency: Judgments on different tasks. *Language Learning* 54 (4), 655–679.
Educational Testing Service (2009) *The Official Guide to the TOEFL Test* (3rd edn). New York: McGraw-Hill.
Field, J. (2005) Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly* 39 (3), 399–423.
Fillmore, C.J. (1979) On fluency. In C.J. Fillmore, D. Kempler and S.-Y. Wang (eds) *Individual Differences in Language Ability and Language Behaviour* (pp. 85–101). New York: Academic Press.
Fulcher, G. (2003) *Testing Second Language Speaking*. Harlow: Longman.
Fulcher, G. (2015) *Re-examining Language Testing: A Philosophical and Social Inquiry*. London and New York: Routledge.
Gass, S. and Varonis, E.M. (1984) The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning* 34 (1), 65–87.
Isaacs, T. and Thomson, R.I. (2013) Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly* 10 (2), 135–159.
Kennedy, S. and Trofimovich, P. (2008) Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review* 64, 459–490.
Kim, Y.-H. (2009) An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing* 26 (2), 187–217.
Koponen, M. and Riggenbach, H. (2000) Overview: Varying perspectives on fluency. In H. Riggenbach (ed.) *Perspectives on Fluency* (pp. 5–24). Ann Arbor, MI: University of Michigan Press.
Lennon, P. (1990) Investigating fluency in EFL: A quantitative approach. *Language Learning* 40 (3), 387–417.
Levelt, W. (1989) *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.
Levelt, W. (1999) Producing spoken language: A blueprint of the speaker. In C. Brown and P. Hagoort (eds) *The Neurocognition of Language* (pp. 83–122). Oxford: Oxford University Press.
Linacre, J.M. (2013) Facets Rasch measurement software (Version 3.71) [computer software]. Chicago, IL: WINSTEPS.com.
Riazantseva, A. (2001) Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition* 23 (4), 497–526.
Riney, T.J., Takagi, N. and Inutsuka, K. (2005) Phonetic parameters and perceptual judgments of accent in English by American and Japanese listeners. *TESOL Quarterly* 39 (3), 441–466.
Rubin, D.L. (1992) Nonlanguage factors affecting undergraduates' judgments of nonnative English speaking teaching assistants. *Research in Higher Education* 33 (4), 511–531.
Schoonen, R. (2012) The generalizability of scores from language tests. In G. Fulcher and F. Davidson (eds) *The Routledge Handbook of Language Testing* (pp. 363–377). London and New York: Routledge.
Segalowitz, N. (2010) *Cognitive Bases of Second Language Fluency*. New York: Routledge.
Smith, L.E. and Nelson, C.L. (1985) International intelligibility of English: Directions and resources. *World Englishes* 4 (3), 333–342.
Van Moere, A. (2012) A psycholinguistic approach to oral language assessment. *Language Testing* 29 (2), 325–344.
Winke, P., Gass, S. and Myford, C. (2011) *The Relationship Between Raters' Prior Language Study and the Evaluation of Foreign Language Speech Samples*. Princeton, NJ: Educational Testing Service.

Winke, P., Gass, S. and Myford, C. (2013) Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing* 30 (2), 231–252.
Xi, X. and Mollaun, P. (2009) How do raters from India perform in scoring the TOEFL iBT speaking section and what kind of training helps? TOEFL iBT Research Report No. RR-09-31. Princeton, NJ: Educational Testing Service.

4 What Can Pronunciation Researchers Learn From Research into Second Language Writing?

Ute Knoch

Introduction

Pronunciation assessment is often high stakes. This is particularly the case for situations where the assessment might determine the legitimacy of the claims of an asylum seeker or in cases where the assessment determines whether an international teaching assistant is allowed to work while undertaking graduate studies. While assessing and teaching second language (L2) pronunciation is therefore clearly important, research in this area is still limited and has only had a recent revival (e.g. Isaacs, 2014). This chapter sets out some possible areas of research and pedagogy in the area of L2 writing that could inform future work on L2 pronunciation. In particular, the chapter focuses on the following areas: (a) rating scale development and validation; (b) rater effects and rater training; (c) task effects; and (d) issues in classroom-based assessment. Where appropriate, the chapter also briefly draws on relevant work that has been done in the area of speaking assessment.

Rating Scale Development and Validation

Rating scale development

Rating scales provide an operational definition of a linguistic construct or a language ability being measured (e.g. Davies *et al.*, 1999), and are interpreted by raters as the de facto test construct (Mcnamara, 2002). Rating scales therefore embody the underlying notion of what abilities are being measured through assessment. In terms of pronunciation assessment, often

little information is available about how rating scales were developed (an issue that is common to most scales in language assessment). For example, the original IELTS sub-scale for pronunciation included only descriptors for four levels out of nine (all even descriptors) until a revision in 2008. Raters were therefore forced to choose between fewer levels. Reports describing the 2008 revision, which was designed to add descriptors for the uneven levels (DeVelle, 2008), do not provide much information about the scale development process apart from the fact that a group of experts (including expert raters) decided on the descriptors for the remaining levels. The new in-between band scales in the revised IELTS scale show, for example, at Band 7 all positive features of Band 6 and some, but not all, of the positive features of Band 8, and similar descriptors appear in Bands 3 and 5. In their IELTS-funded research projects, Yates *et al.* (2011) and Isaacs *et al.* (2015) documented that these descriptors are problematic and that IELTS examiners find them at times difficult to work with. In another pronunciation study, Isaacs and Thomson (2013) developed two different sets of scales for comprehensibility, fluency and accentedness, with one scale for each construct featuring five levels and the other, nine levels. Following conventions in SLA-oriented L2 pronunciation research (e.g. Derwing & Munro, 1997; Munro & Derwing, 1995), Isaacs and Thomson chose numerical scales with only the end-points labelled. For example, the scale for accentedness ranged from 'heavily accented' to 'not accented at all'. The two examples of studies targeting scale development or validation activities described above both developed scales for pronunciation in different ways. In one example, researchers used a group of experts, whereas in the other they chose non-labelled scales following research conventions.

Researchers in the area of L2 writing assessment, for which there is a longer tradition of scale development, have described a number of considerations when developing a rating scale (e.g. Weigle, 2002). These include the type of rating scale to use as well as what to base the criteria on. Two main types of rating scales (e.g. holistic and analytic scales) have been used to assess writing performances, and these formats are also suitable for the assessment of L2 pronunciation. Holistic scales provide overall descriptions of performance at different score levels while analytic scales break these performances down into different criteria. If pronunciation is one aspect of many in a speaking assessment (such as the IELTS speaking scale), it may be included as one criterion in an analytic scale. If pronunciation is the only object of measurement (e.g. for a pronunciation test or a research study), then developers need to decide whether a holistic scale or an analytic scale is most suitable.

In order to decide what weight pronunciation is given in an assessment, it is important to take the targeted construct into account, as well as what decisions the assessment is designed to support. It is clear that the construct of pronunciation encompasses a number of distinct sub-areas such as, among

others, rhythm, stress and individual sounds. Depending on the context, it might therefore be worth using an analytic scale for a finer grained evaluation of different features of pronunciation. At the same time, it is not always easy to separate pronunciation from other areas of speech (e.g. accuracy, interactional features), and a similar problem exists with features of writing (e.g. separating grammatical accuracy from complexity). The purpose of the assessment also influences the scale type chosen. For example, for a diagnostic assessment whose goal is to provide feedback on strengths and weaknesses, the scale chosen would either be analytic or even more detailed, for example, in the form of a checklist.

One of the main considerations in scale development needs to be what the criteria and descriptors are based on. This is important, as the rating scale acts as a representation of the construct of an assessment. Although usually not much information is available on scale development methods (but see North, 1998), the most common method of development is often based on intuitions rather than an explicit theoretical model (Fulcher, 2003). These intuitions can come from one expert or a committee of experts (such as in the case of the pronunciation scale for IELTS) and may be experiential, in that a scale goes through several iterations of changes and possibly starts from a previously existing scale and is adapted for a new context by experts. However, these development methods have been criticized as they often lack a theoretical basis and might be under-representing or at times even misrepresenting the construct.

Knoch (2011b) reviewed several possible models or theories that can form the basis for a writing assessment. For example, models of communicative competence (Bachman, 1990; Bachman & Palmer, 1996) have been used as the basis for rating scales for writing. However, these may not be detailed enough for the assessment of pronunciation (Isaacs, 2014) because these models have been developed to account for many contexts and situations of language learning and use, and pronunciation is usually only a sub-construct of linguistic competence. For writing, models of text construction (e.g. Grabe & Kaplan, 1996) were also reviewed; Knoch concluded that these models are not sufficiently developed as yet to form the basis of a rating scale. She therefore recommended using a combination of several theories and models as the basis for identifying the criteria to include in a scale. This makes it possible to include measureable aspects of language into a rating scale while at the same time excluding aspects of theories or models that cannot easily be operationalized into a scale of language proficiency (e.g. the use of background knowledge). This may also be an option for test developers and researchers working on the development of a scale for pronunciation.

Other researchers have developed their scales empirically, basing their descriptions on observable, empirically verifiable learner data (e.g. using detailed descriptions or listener-coded measures). It is generally understood that empirical scale development methods avoid problems often reported in

a priori rating scale development. These problems involve features in the scales which are not seen in actual performances or terminology in descriptors being unnecessarily subjective (e.g. Brindley, 1998; Mickan, 2003; Turner & Upshur, 2002; Upshur & Turner, 1995). Empirically driven scale development has been used in the development of assessment instruments for fluency (e.g. Fulcher, 1987, 1996) and comprehensibility (Isaacs & Trofimovich, 2012), in addition to general speaking scales (described further below; Turner & Upshur, 2002; Upshur & Turner, 1995), including instruments linked to assessment for specific purposes (e.g. health care or the tourist industry, see Elder *et al.*, 2013; Fulcher *et al.*, 2011). Going beyond the research methods described above, work on L2 writing has additionally drawn on corpus linguistic techniques (Hawkey, 2001; Hawkey & Barker, 2004) to identify salient features at pre-scored writing levels so that these features could be included in the eventual writing scale.

A further empirical scale development method (empirically derived, binary-choice, boundary definition scales, or EBBs) was first proposed by Upshur and Turner (1995, 1999). This method requires a group of experienced teachers or raters to group performances into separate piles based on raters' perceptions of writers' relative ability, and then to identify the distinguishing features at each level. Once the scales are developed, raters make a number of yes/no choices at each level of the flowchart-like scale to arrive at a final score. This method has been used for both writing and speaking but may be suitable for pronunciation assessment if more fine-grained decisions are necessary.

Exemplifying another method of scale development for writing, Knoch (2007, 2009) combined empirical methods with a review of theoretical frameworks, as described above. She first identified the criteria used in several relevant theories and models of writing and then derived rating scale descriptors by carefully analyzing a large number of writing scripts at different levels (i.e. empirically). Her findings resulted in defined descriptors at different levels, which were identified through a detailed discourse analysis followed by a statistical analysis. The resulting scale was validated using a many-facet Rasch analysis which is powerful in its detailed analysis of rating scale properties (Eckes, 2011; Knoch & Mcnamara, 2015).

Methods such as the ones described above may well be extended to pronunciation assessment. However, it is important to remember that, in the case of many assessments of speaking, decisions are time bound, which means that raters make decisions in real time (and often act as interviewers or interlocutors as well), whereas writing raters are often less likely to rely on their memory when making the rating. This is likely to lead to a reduced cognitive load for raters assessing writing, compared to those evaluating speaking (Knoch, 2009). In such situations where live performances are rated, innovative scale development methods could be explored which require raters to make some initial broad distinctions while finer, more detailed

decisions are left until after the end of the performance (see also Saito *et al.*, this volume, for assessment of speaking through rating transcripts of L2 speech). Scales such as the EBBs described above may be useful in such contexts. It is important to note that empirically developed rating scales are usually task specific, which means that an EBB derived from a read-aloud task may be different from an instrument developed for a spontaneous speaking task. Scale generalizability may therefore be an issue to consider.

Rating scale validation

Methods of scale validation in language assessment have historically followed more general conceptualizations of validity in the educational literature (e.g. Chapelle, 2012). More recently, argument-based approaches to validity (Bachman & Palmer, 2010; Chapelle *et al.*, 2008; Kane, 1992, 2006) have become prominent in the area of language assessment. These frameworks provide a structured way to examine multiple aspects of an assessment and the use of scores. These include an examination of the conditions under which a task is administered, task characteristics, rating scales and rating procedures, as well as assessment reliability. These also involve an analysis of the construct measured by the assessment, the relevance of what the assessment measured to the wider world from which the language is sampled, and research on the utility of the results to the users. While researchers working on rating scale validation have drawn on a variety of methods to collect empirical evidence, they have usually not focused on a specific framework to guide their validation work. This pattern is evident in work on rating scale validation for writing, where researchers have most commonly used statistical techniques to show that a scale can be used reliably (e.g. East, 2009; Harsch & Martin, 2012), how well different scale categories (levels) perform (e.g. Knoch, 2007, 2009), and how many different dimensions subcategories of an analytic scale measure (e.g. Knoch, 2009; Zhao, 2013). Others have complemented these quantitative methods with more qualitative analyses, which included interviews or group discussions with raters (e.g. Harsch & Martin, 2012), think-aloud protocols of raters applying the scale (e.g. DiPardo *et al.*, 2011) and, on rare occasions, student attitudes towards assessment criteria (e.g. Morozov, 2011). A mixed-methods approach combining both quantitative and qualitative techniques for rating scale validation is powerful as it can draw out issues that may be masked by using a single method (e.g. Harsch & Martin, 2012).

Work on rating scale validation in the area of pronunciation assessment (e.g. Harding, this volume; Isaacs & Thomson, 2013; Yates *et al.*, 2011) has been more scarce, compared to research in writing assessment. As discussed previously, Isaacs and Thomson compared rating scales of varying lengths and undertook detailed analyses to explore scale functioning by examining scale response category plots. Yates *et al.* (2011), on the other hand, focused

mainly on qualitative responses from raters collected through questionnaires (although some basic statistical analyses were carried out using questionnaire responses) and think-aloud protocols collected from a small number of raters using the scale. Neither study made use of a specific validation framework.

The brief review of rating scale validation efforts presented above shows that researchers and test developers interested in validating rating scales typically draw on a number of methods, both qualitative and quantitative. However, there is usually no specific test validation framework guiding researchers and test developers. Part of the reason for this is that a framework of this type is not readily available in the literature. Knoch (2007, 2009), for example, adapted Bachman and Palmer's (1996) test usefulness framework to guide the validation efforts for her rating scale for diagnostic assessment. More recently, Knoch and Chapelle (in preparation) have been working on integrating rating processes into an argument-based framework of validation in order to more closely set out the areas researchers should focus on in the validation process. This is important as these aspects have often been underspecified. They show that research on issues relating to rating scales is important across a range of areas of validation research in our field. For example, researchers can investigate not only the reliability of scales, how well the scale steps (levels) are functioning and how well the scale steps are able to distinguish between test taker ability levels, but can also focus their investigations on how well the writing construct is captured in the scale and how relevant the scale is to the wider domain being measured. Researchers can also examine whether the scale has positive consequences on stakeholders. Some examples of such work might focus on:

- a review of the construct coverage of the scale;
- how well raters' cognitive processes are in line with the theoretical model the scale is based on;
- how well discourse produced in response to a certain task is reflective of the rating scale descriptions;
- whether the scale criteria adequately reflect the evaluation criteria of the real-world context;
- whether the scale layout is appropriate for score reporting and decision making by test users;
- whether test users are able to interpret the scale criteria; and
- whether the scale has a positive impact on teaching and learning.

It is clear from this information that scale validation efforts have often been only narrowly applied to investigating issues of reliability and more work in this area is certainly needed, in particular in the area of validation of scales for pronunciation.

Rater Effects and Training

Rater effects have been shown in research targeting both writing and speaking assessment. This is an important area to investigate, as rater effects may impact the outcomes of ratings and introduce construct-irrelevant variance, that is, extraneous 'noise' which may influence scores without being relevant to the measurement. In the following section, I examine some of the research on rater effects and efforts in rater training to combat such unwanted influences and draw parallels to work in L2 pronunciation research.

Rater effects

The impact of rater effects on writing scores has been studied extensively. Most commonly, studies have employed verbal protocols to identify any specific rating behaviours during the rating process or statistical methods, such as multi-faceted Rasch analyses, to generate bias interaction reports between raters' behaviour and specific elements of the rating situation (e.g. rating scale criteria, task types or test taker background characteristics). In a much-cited study by Vaughan (1991), for example, verbal protocols were collected from raters during the rating process, and a number of distinct reading styles were identified in a small group of raters. It was concluded that these individual styles might introduce construct-irrelevant variance and that a clearer understanding of these styles was necessary. More recently, Schaefer (2008) examined bias interactions of relatively inexperienced raters with analytic scoring criteria using multi-faceted Rasch analysis and was able to sort raters into groups of rating behaviour. However, these studies failed to identify specific reasons for rater biases or rating behaviours.

Other studies have investigated the influence of predetermined rater background variables on test scoring. Typical variables examined in this type of research are rater background variables, such as the distinction between novice and experienced raters, raters' disciplinary background, raters' own language or language learning background, or general decision-making style (Baker, 2012; Barkaoui, 2010; Cumming, 1990; Johnson & Lim, 2009; Xi & Mollaun, 2009). Many of these background variables have been found to introduce construct-irrelevant variance into the rating process and therefore to influence scores. Researchers have therefore called for more careful rater monitoring and awareness-raising during rater training.

Combining the two research areas described above, Eckes (2008) advanced a rater type hypothesis and was able to show, employing a large dataset, that the relative importance raters pay to different criteria as well as a number of rater background variables resulted in six distinct rater types. He argued that a deeper understanding of rater background, personal preferences and cognitive processing has a significant impact on score meaning and interpretation and should be addressed operationally and in rater training.

Research into rater effects is of course not specific to writing assessment, although fewer studies seem to have focused on speaking assessment. As was the case with writing, research has examined rater bias towards certain scale criteria and groups of test takers from the same language background (e.g. Yan, 2014), without examining possible causes of such biases. Other studies have focused on examining specific rater background variables and have investigated whether these can be attributed to certain biases in a dataset. Among the background variables under investigation have been rater experience (e.g. Brown, 2000), native language (L1) background of raters and bias displayed with regard to the weighting of specific criteria (e.g. Kim, 2009), or rating behaviour in general (e.g. Zhang & Elder, 2011). One study, targeting the link between rater and examinee characteristics, has examined the interaction between raters' L2 and test takers' L1 background (Winke *et al.*, 2013). It was found that groups of raters familiar with the L1 of students (through personal study) favoured those L1 groups.

Finally, Isaacs and Thomson (2013) used verbal report methodology to show how experienced and novice raters draw on their experience (in the case of experienced raters) or offset their lack of experience (in the case of novice raters) in the rating of pronunciation. More specifically, raters with less experience lacked the terminology to express pronunciation features that they may have been attending to. Isaacs and Trofimovich (2010) also found that musically trained raters tend to be more severe in their comprehensibility scoring than raters with less training in music, although the findings were tentative and their implications for high-stakes assessments have yet to be explored.

From the review presented above, it is clear that studies examining the influence of specific rater background variables, rater orientations and rater decision-making styles on one or more pronunciation scale criteria (e.g. comprehensibility, goodness of prosody) are needed. As pronunciation features are often salient and provide raters with an indication of a test taker's L1 background, it is important to examine in more detail what influence rater characteristics such as L1 background, L2 experience, experience teaching students from certain L1 backgrounds, or attitudes to accents may have on raters' assessment of pronunciation and on their evaluation of specific pronunciation sub-areas, such as prosody or fluency. More work is certainly required in this area.

Rater training

Acknowledging the fact that raters differ in their ratings, a number of studies in the area of writing assessment have set out to investigate whether rater training can reduce some of the between-rater differences. Such training usually involves groups of raters being (re-)introduced to the assessment criteria, and raters marking a number of benchmark performances chosen to

represent a variety of proficiency levels and typical problem scripts. Studies examining the effects of group rater training (e.g. Weigle, 1994, 1998) have shown that such training can be effective in that it eliminates extreme differences in rater severity, increases the self-consistency of raters and reduces individual biases displayed by raters towards certain aspects of the rating situation. However, many researchers have acknowledged that differences between raters continue to exist despite such training sessions. One attempt to address these differences has been made through providing individual feedback to raters (e.g. Elder *et al.*, 2005; Knoch, 2011a; O'Sullivan & Rignall, 2007). In these studies, raters were provided with feedback on their relative severity in relation to the group of raters, on their consistency when rating, and on any individual biases displayed in relation to any rating scale criteria. Training effectiveness varied across studies, with some studies reporting more successful training than others. This might be due to different methodologies employed (e.g. whether control groups were tested, whether feedback was provided longitudinally or in one-shot designs). Overall, these studies show that the relative severity of raters' decisions can be treated more easily with feedback, compared to rater inconsistencies.

Given the findings of studies such as the one conducted by Yates *et al.* (2011), which highlight some of the difficulties raters experience when using pronunciation rubrics (i.e. IELTS pronunciation sub-scale), it may be worthwhile examining rater training for pronunciation in more detail. As is the case with certain criteria in writing assessment (e.g. coherence, cohesion), raters' understanding of specific features of pronunciation may need extra training (including accessible self-training documentation). Raters may also need regular feedback so that they can become intimately familiar with rating criteria and can understand possible personal biases and attitudes towards certain accents or specific speech patterns. More research in this area is clearly needed.

Task Effects

Researchers working on writing and writing assessment often conduct detailed analyses of test takers' discourse while performing particular test tasks in order to gain a clear understanding of the language produced by test takers. In the area of writing assessment, this is a frequent focus of validation work because such analyses can, for example, provide support for the descriptors in a rating scale by showing that there are differences in test takers' language across score levels. This analysis can then feed back into refinements of rating scale descriptors and its findings can also inform scoring algorithms used in automated scoring programs. More recently, with integrated tasks becoming more popular (e.g. Gebril & Plakans, 2013; Plakans, 2009), it is also important for test providers to show that different

tasks in a speaking test in fact elicit different discourse types from test takers. If no differences were found, test developers would lack evidence for including extra tasks. The same type of work can also provide evidence for the use of the same or task-specific rating scales for rating.

Several studies in the area of writing assessment have included detailed discourse analyses to examine whether different tasks in fact elicit different language from test takers. Two of the most cited studies were conducted in the context of the Test of English as a Foreign Language (TOEFL). Cumming *et al.* (2006), for example, carried out a study on the TOEFL new generation prototype tasks (an independent task, a reading to write task, and a listening to write task), documenting an effect for task type. These researchers found that students wrote shorter essays, used longer words and a wider range of words, wrote more and longer clauses, and wrote less argumentatively oriented texts when responding to integrated than to independent tasks. More recently, Knoch *et al.* (2014) investigated differences in discourse between the two TOEFL writing tasks that were adopted in the new TOEFL (independent task versus integrated listening and reading task). They found that test takers wrote significantly more words, clauses and *t*-units, more lexically rich essays, and used a higher proportion of self-voice when responding to the independent than to the integrated task. On the other hand, the essays written in response to the integrated task displayed significantly longer words, higher lexical density and lexical sophistication, and higher levels of coherence and density of anaphora. The proportion of references to others as well as the percentage of material copied from the input (mainly the reading) was also higher in responses to the integrated task, compared to the independent task.

Similar work has also been conducted in the area of speaking assessment. In one study, again focusing on the TOEFL, Brown *et al.* (2005) examined similar effects across task types for one integrated and one independent speaking task. They employed a number of measures of pronunciation in their study; however, they did not include more than one form of each task type and only included two task types. Their study therefore needs to be replicated so that their validity can be confirmed. A more recent study by Biber and Gray (2013), comparing performances on TOEFL speaking (and writing) tasks using corpus-linguistic techniques, focused on lexicogrammatical measures only and therefore did not target any measures related to pronunciation. Saito *et al.* (2016) have started to address this by looking at comprehensibility and accentedness as a function of speakers' ability levels, using a number of measures for native Japanese and French speakers of English.

There is clear need for more work in this area. For example, it would be important to investigate whether there are any differences in pronunciation features produced by test takers responding to semi-direct tasks such as the ones employed by the TOEFL, and tasks requiring interaction with an

examiner (e.g. Crowther *et al.*, 2015), another test taker or a group of test takers. In the case of paired or group assessments, it would be interesting to investigate what influence pronunciation features of the interlocutors have on the speech of test takers and on the pronunciation scores they receive.

Classroom-based Assessment

Diagnostic assessment

Diagnostic assessment, which is designed to identify the specific strengths and weaknesses of learners in certain aspects of language, is underresearched and has only recently experienced a revival due to a book published by Alderson (2005). While a number of assessments are called diagnostic, if closely scrutinized many of these are likely not truly diagnostic. Alderson (2005) set out a list of principles or criteria for diagnostic tests which they should fulfil, regardless of the skills such assessments are designed to assess. According to Alderson, diagnostic assessments should be based on either a theory of language or language development, or should be closely linked to a curriculum. They should focus on specific strengths and weaknesses of learners and on content that is understandable to students, and can be closely linked to future learning and teaching. Diagnostic feedback should be broken down into accessible chunks and should come with a recommendation that learners can act upon.

Knoch (2007, 2009) developed a rating scale for diagnostic writing assessment for a large-scale university post-entry assessment. Although her scale was developed to be used by assessors, she argued that the feedback to students could be presented with accessible explanations and examples associated with each scoring level, as well as with clear recommendations as to where on campus students could find support for their specific weaknesses. In a more recent study, Wagner (2014) developed a detailed checklist for the diagnostic assessment of high school students' writing (i.e. for opinion-based essays) in Ontario schools. Her diagnostic rubric for assessing writing is grounded in theory and is curriculum relevant; it also focuses on students' use of writing knowledge, skills and strategies rather than on discrete knowledge (such as certain error categories) and is designed so it could also be adapted for self-assessment.

Test developers and researchers interested in developing a diagnostic scale for pronunciation assessment need to identify the sub-skills or aspects of pronunciation they would like to include in the scale (e.g. pronunciation of vowels and consonants, intonation, rhythm, stress, etc.). These could each be expressed as a checklist or as a scale (e.g. North, 2003). For the diagnostic assessment to have maximum impact in terms of future teaching and learning, the terms used in the scale should be accessible to students. It is therefore

important that training is given to students and that clear examples are included. It is also recommended that clear recommendations are provided on where students can access help and materials to work on potential weaknesses identified through testing. Apart from the scale used in diagnostic assessment, it is also important to choose an appropriate task that is meaningful to students and that elicits the specific features targeted in the scale or checklist. Follow-up research should focus on students' use of the feedback and recommendations as well as on teachers' integration of the results into their lesson planning and teaching.

Peer assessment

Research into peer assessment has focused on examining the effectiveness of peer assessment, including comparisons of peer and teacher feedback. These studies have shown that, when provided with both types of feedback, students are generally more likely to incorporate teachers' comments than comments from peers, and that the effectiveness of peer feedback varies across studies (e.g. Connor & Asenavage, 1994; Paulus, 1999). Some advantages of peer feedback over teacher feedback have also been identified (apart from the obvious time saving for teachers). Zhao (2010), for example, found that peers were more likely to understand the comments from peers, and that providing feedback also raised students' awareness of their own writing problems. Peer feedback may also impact on more meaning-level errors, whereas teacher feedback may impact on more surface-level errors (e.g. Yang *et al.*, 2006).

Research on peer assessment has also focused on examining factors that may affect the efficacy of peer feedback. One factor that has been examined is the nature of the interaction in peer feedback dyads (for example, Nelson & Murphy, 1993 made the distinction between co-operative dyads and defensive dyads). A second factor relates to learners' proficiency in peer feedback dyads. Kamimura (2006), for instance, compared two intact classes in terms of the benefits of peer feedback, one of low-proficiency students and one of high-proficiency students. Compared to students in the low-proficiency class, students in the high-proficiency class improved more. In addition, the feedback provided in the high-proficiency class focused more on global-level errors, while their lower proficiency counterparts focused more on surface-level errors. Kamimura did not, however, compare mixed dyads.

A final and crucial aspect of peer feedback that has also been investigated is the role of the training of feedback providers (and receivers). Training in how to properly engage in peer feedback sessions is crucial to making peer feedback work. Interestingly, the level of training has varied greatly from study to study, ranging from no training or very short training (approximately 30 minutes) to approximately seven hours (e.g. Berg, 1999; Min, 2005; Rahimi, 2013; Stanley, 1992). The training has also varied greatly in focus, in terms of guidelines and/or checklists provided to students. These

aspects of training will undoubtedly have an influence on the effectiveness of the feedback provided.

Peer assessment of pronunciation is an area that is underexplored in the research literature and certainly an area in need of research. Some studies employing peer feedback in reading aloud activities (e.g. Tost, 2013) have found a positive effect on pronunciation, but research – in line with comparable literature on writing – overall seems scarce. Providing peers with training in the use of specifically designed peer feedback checklists, especially in mixed L1 classes, could be an effective method of pronunciation assessment and could supplement speaking activities in classrooms. Similarly, a close examination of pre- and post-test results could examine what types of interactions in dyads are the most effective or whether the students who have travelled more widely, and have therefore been exposed to a wider range of accents or pronunciation features, are more likely to be able to provide feedback on a range of pronunciation features. It would also be interesting to examine whether students are able to identify pronunciation errors in their peers' speech and whether they are able to pinpoint specific problems. Finally, it would be of interest to examine whether there are positive effects of peer feedback for the pronunciation of the feedback provider, in line with what has been shown in writing assessment (Lundstrom & Baker, 2009).

Implications and Conclusion

As described above, the chapter draws out a number of implications for research on L2 pronunciation assessment and professional practice. For example, in the area of rating scale development and validation, future research could focus on examining the validity of scale descriptors being developed by drawing both on theoretical perspectives on L2 pronunciation development as well as actual data from student performances. Research in the area of rating scale validation can also be expanded beyond a narrow focus on scale category functioning and rater feedback, in order to examine more closely verbal thought processes when rating and how well they align with the theoretical construct of the pronunciation assessment, whether scale layout is suitable for user decision-making purposes, how well test users can understand the scale, and whether the scale has an impact on teaching and learning. Further research could also examine what influence specific rater characteristics, rater orientations and rater decision-making styles have on test takers' pronunciation scores.

In terms of pedagogical and practical implications, further work could focus on examining the effectiveness of the use of diagnostic descriptors for pronunciation in classroom contexts. Similarly, more research is needed to explore whether peer feedback on pronunciation is effective. Practitioners

engaged in L2 assessment may also want to investigate the effectiveness of rater training on pronunciation ratings. The same line of enquiry could explore providing feedback to pronunciation raters and the effectiveness of training raters in online environments.

This chapter has drawn on research and scholarship in L2 writing to describe current developments in a number of areas, including rating scale development and validation, rater and task effects, and classroom-based assessment. Parallels were drawn to areas in which L2 pronunciation pedagogy and assessment could draw on some of these findings to expand current practice and broaden existing research agendas. While similar work has been undertaken in the area of L2 speaking and speaking assessment, work in L2 writing might help inform current practice in L2 pronunciation.

References

Alderson, C. (2005) *Diagnosing Foreign Language Proficiency: The Interface Between Learning and Assessment*. London: Continuum.

Bachman, L.F. (1990) *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. and Palmer, A.S. (1996) *Language Testing in Practice*. Oxford: Oxford University Press.

Bachman, L. and Palmer, A.S. (2010) *Language Assessment in Practice*. Oxford: Oxford University Press.

Baker, B.A. (2012) Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly* 9, 225–248.

Barkaoui, K. (2010) Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly* 7, 54–74.

Berg, E.C. (1999) The effects of trained peer response on ESL students' revision types and writing quality. *Journal of Second Language Writing* 8, 215–241.

Biber, D. and Gray, B. (2013) Discourse characteristics of writing and speaking tasks types on the TOEFL iBT Test: A lexico-grammatical analysis. TOEFL Research Report No. RR-13-04. Princeton, NJ: Educational Testing Service.

Brindley, G. (1998) Describing language development? Rating scales and SLA. In L.F. Bachman and A.D. Cohen (eds) *Interfaces Between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press.

Brown, A. (2000) An investigation of the rating process in the IELTS oral interview. In J. Osborne (ed.) *IELTS Research Reports* (pp. 49–84). Canberra and Manchester: IDP IELTS Australia and British Council.

Brown, A., Iwashita, N. and Mcnamara, T. (2005) *An Examination of Rater Orientation and Test-taker Performance on English-for-academic-purposes Speaking Tasks*. TOEFL Monograph Series No. MS-29. Princeton, NJ: Educational Testing Service.

Chapelle, C. (2012) Conceptions of validity. In G. Fulcher and F. Davidson (eds) *The Routledge Handbook of Language Testing* (pp. 21–33). New York: Routledge.

Chapelle, C., Enright, M. and Jamieson, J. (eds) (2008) *Building a Validity Argument for the Test of English as a Foreign Language*. New York: Routledge.

Connor, U. and Asenavage, K. (1994) Peer response groups in ESL writing classes: How much impact on revision? *Journal of Second Language Writing* 3, 257–276.

Crowther, D., Trofimovich, P., Isaacs, T. and Saito, K. (2015) Does speaking task affect second language comprehensibility? *Modern Language Journal* 99, 80–95.

Cumming, A. (1990) Expertise in evaluating second language compositions. *Language Testing* 7, 31–51.
Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U. and James, M. (2006) *Analysis of Discourse Features and Verification of Scoring Levels for Independent and Integrated Prototype Writing Tasks for the New TOEFL*. TOEFL Monograph Series No. MS-30. Princeton, NJ: Educational Testing Service.
Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and Mcnamara, T. (1999) *Dictionary of Language Testing*. Cambridge: Cambridge University Press.
Derwing, T.M. and Munro, M.J. (1997) Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition* 20, 1–16.
DeVelle, S. (2008) The revised IELTS pronunciation scale. *Research Notes* 34, 36–38.
DiPardo, A., Storms, B.A. and Selland, M. (2011) Seeing voices: Assessing writerly stance in the NWP Analytic Writing Continuum. *Assessing Writing* 16, 170–188.
East, M. (2009) Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing* 14, 88–115.
Eckes, T. (2008) Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing* 25, 155–185.
Eckes, T. (2011) *Introduction to Many-Facet Rasch Measurement*. Frankfurt: Peter Lang.
Elder, C., Knoch, U., Barkhuizen, G. and von Randow, J. (2005) Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly* 2, 175–196.
Elder, C., Mcnamara, T., Woodward-Kron, R., Manias, E., McColl, G. and Webb, G. (2013) Towards improved healthcare communication. Development and validation of language proficiency standards for non-native English speaking health professionals. *Final Report for the Occupational English Test Centre*. Melbourne: University of Melbourne.
Fulcher, G. (1987) Tests of oral performance: The need for data-based criteria. *ELT Journal* 41, 287–291.
Fulcher, G. (1996) Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13, 208–238.
Fulcher, G. (2003) *Testing Second Language Speaking*. London: Pearson Longman.
Fulcher, G., Davidson, F. and Kemp, J. (2011) Effective rating scale development for speaking tests: Performance decision trees. *Language Testing* 28, 5–29.
Gebril, A. and Plakans, L. (2013) Toward a transparent construct of reading-to-write tasks: The relationship between discourse features and proficiency. *Language Assessment Quarterly* 10, 9–27.
Grabe, W. and Kaplan, R.B. (1996) *Theory and Practice of Writing*. New York: Longman.
Harsch, C. and Martin, G. (2012) Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing* 17, 228–250.
Hawkey, R. (2001) Towards a common scale to describe L2 writing performance. *Cambridge Research Notes* 5, 9–13.
Hawkey, R. and Barker, F. (2004) Developing a common scale for the assessment of writing. *Assessing Writing* 9, 122–159.
Isaacs, T. (2014) Assessing pronunciation. In A.J. Kunnan (ed.) *The Companion to Language Assessment* (pp. 140–155). New York: John Wiley.
Isaacs, T. and Thomson, R.I. (2013) Rater experience, rating scale length, and judgements of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly* 10, 135–159.
Isaacs, T. and Trofimovich, P. (2010) Falling on sensitive ears? The influence of musical ability on extreme raters' judgements of L2 pronunciation. *TESOL Quarterly* 44, 375–386.
Isaacs, T. and Trofimovich, P. (2012) 'Deconstructing' comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition* 34, 475–505.

Isaacs, T., Trofimovich, P., Yu, G. and Chereau, B.M. (2015) Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS Research Reports Online* 4, 1–48.

Johnson, J.S. and Lim, G.S. (2009) The influence of rater language background on writing performance assessment. *Language Testing* 26, 485–505.

Kamimura, T. (2006) Effects of peer feedback on EFL student writers at different levels of English proficiency: A Japanese context. *TESL Canada Journal* 23, 12–39.

Kane, M. (1992) An argument-based approach to validity. *Psychological Bulletin* 112, 527–535.

Kane, M. (2006) Validation. In R.L. Brennan (ed.) *Educational Measurement* (pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kim, Y.-H. (2009) An investigation into native and non-native teachers' judgements of oral English performance: A mixed methods approach. *Language Testing* 26 (2), 187–217.

Knoch, U. (2007) The development and validation of an empirically-developed rating scale for academic writing. Unpublished PhD dissertation, University of Auckland.

Knoch, U. (2009) *Diagnostic Assessment of Writing: The Development and Validation of a Rating Scale*. Frankfurt: Peter Lang.

Knoch, U. (2011a) Investigating the effectiveness of individualized feedback to rating behaviour: A longitudinal study. *Language Testing* 28, 179–200.

Knoch, U. (2011b) Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing* 16, 81–96.

Knoch, U. and Chapelle, C. (in preparation) Validation of rating processes within an argument-based framework.

Knoch, U. and Mcnamara, T. (2015) Rasch analysis. In L. Plonsky (ed.) *Advancing Quantitative Methods in Second Language Research* (pp. 275–304). New York: Taylor and Francis/Routledge.

Knoch, U., Macqueen, S. and O'Hagan, S. (2014) An investigation of the effect of task type on the discourse produced by students at various score levels in the TOEFL iBT writing test. TOEFL iBT Report No. 23; ETS Research Report No. RR-14-43. Princeton, NJ: Educational Testing Service.

Lundstrom, K. and Baker, W. (2009) To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing* 18, 30–43.

McNamara, T. (2002) Discourse and assessment. *Annual Review of Applied Linguistics* 22, 221–242.

Mickan, P. (2003) 'What's your score?' An investigation into language descriptors for rating written performance. In J. Osborne (ed.) *IELTS Research Reports* (pp. 126–155). Manchester/Canberra: IDP IELTS Australia and British Council.

Min, H.-T. (2005) Training students to become successful peer reviewers. *System* 33, 293–308.

Morozov, A. (2011) Student attitudes toward the assessment criteria in writing-intensive college courses. *Assessing Writing* 16, 6–31.

Munro, M.J. and Derwing, T.M. (1995) Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning* 45, 73–97.

Nelson, G.L. and Murphy, J.M. (1993) Peer response groups: Do L2 writers use peer comments in revising their drafts? *TESOL Quarterly* 27, 113–131.

North, B. (1998) Scaling descriptors for language proficiency scales. *Language Testing* 15, 217–263.

North, B. (2003) *Scales for Rating Language Performance: Descriptive Models, Formulation Styles, and Presentation Formats*. TOEFL Monograph Series No. MS-24. Princeton, NJ: Educational Testing Service.

O'Sullivan, B. and Rignall, M. (2007) Assessing the value of bias analysis feedback to raters for the IELTS writing module. In L. Taylor and P. Falvey (eds) *IELTS Collected Papers* (pp. 446–478). Cambridge: Cambridge University Press.

Paulus, T.M. (1999) The effect of peer and teacher feedback on student writing. *Journal of Second Language Writing* 8, 265–289.

Plakans, L. (2009) Discourse synthesis in integrated second language writing assessment. *Language Testing* 26, 561–587.

Rahimi, M. (2013) Is training student reviewers worth its while? A study of how training influences the quality of students' feedback and writing. *Language Teaching Research* 17, 67–89.

Saito, K., Trofimovich, P. and Isaacs, T. (2016) Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics* 37 (2), 217–240.

Schaefer, E. (2008) Rater bias patterns in an EFL written assessment. *Language Testing* 25, 465–493.

Stanley, J. (1992) Coaching student writers to be effective peer evaluators. *Journal of Second Language Writing* 1, 217–233.

Tost, G. (2013) Bettering pronunciation through reading aloud and peer appraisal. *Bellaterra Journal of Teaching and Learning Language and Literature* 6, 1–15.

Turner, C.E. and Upshur, J.A. (2002) Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly* 36, 49–70.

Upshur, J.A. and Turner, C.E. (1995) Constructing rating scales for second language tests. *ELT Journal* 49, 3–12.

Upshur, J.A. and Turner, C.E. (1999) Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing* 16, 82–111.

Vaughan, C. (1991) Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (ed.) *Assessing Second Language Writing in Academic Contexts*. Norwood, NJ: Ablex.

Wagner, M. (2014) Use of a diagnostic rubric for assessing writing: Students' perceptions of cognitive diagnostic feedback. Paper presented at the Language Testing Research Colloquium, Amsterdam, June.

Weigle, S.C. (1994) Effects of training on raters of ESL compositions. *Language Testing* 11, 197–223.

Weigle, S.C. (1998) Using FACETS to model rater training effects. *Language Testing* 15, 263–287.

Weigle, S.C. (2002) *Assessing Writing*. Cambridge: Cambridge University Press.

Winke, P., Gass, S. and Myford, C. (2013) Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing* 30, 231–252.

Xi, X. and Mollaun, P. (2009) How do raters from India perform in scoring the TOEFL iBT speaking section and what kind of training helps? TOEFL iBT Research Report No. RR-09-31. Princeton, NJ: Educational Testing Service.

Yan, X. (2014) An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing* 31, 501–527.

Yang, M., Badger, R. and Yu, Z. (2006) A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing* 15, 179–200.

Yates, L., Zielinski, B. and Pryor, E. (2011) The assessment of pronunciation and the new IELTS Pronunciation Scale. In J. Osborne (ed.) *IELTS Research Reports* 12 (pp. 1–46). Melbourne and Manchester: IDP IELTS Australia and British Council.

Zhang, Y. and Elder, C. (2011) Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing* 28, 31–50.

Zhao, C.G. (2013) Measuring authorial voice strength in L2 argumentative writing: The development and validation of an analytic rubric. *Language Testing* 30, 201–230.
Zhao, H. (2010) Investigating learners' use and understanding of peer and teacher feedback on writing: A comparative study in a Chinese English writing classroom. *Assessing Writing* 15, 3–17.

5 The Role of Pronunciation in the Assessment of Second Language Listening Ability

Elvis Wagner and Paul D. Toth

Introduction

Traditionally, second language (L2) pronunciation has been operationalized as a component of speaking ability, and efforts to assess L2 learners' pronunciation have focused on the test taker's spoken production. But pronunciation also plays a role in the assessment of L2 listening ability. L2 listening tests almost invariably utilize recorded spoken texts to assess the test taker's comprehension. Isaacs (2014) argues that it is essential in L2 speaking assessment to define and clarify the role of pronunciation within this construct. However, we believe that it is also necessary to do so in the assessment of L2 listening, because how the speakers of L2 texts articulate their utterances can impact on comprehensibility and, consequently, test taker performance. This chapter explores a number of points related to this issue, including: how the spoken texts used in L2 listening tests are chosen; the effects of scripted versus unscripted texts; the organization, phonology and fluency characteristics of spoken texts; and how these issues impact on construct validity. The chapter then presents an empirical study investigating L2 test takers' beliefs about the nature of spoken texts used in an L2 listening test.

Review of the Literature

Spoken texts used in L2 tests

In theory, the target language use (TLU) domain of interest dictates the types of spoken texts that will be used in L2 listening assessments (Bachman & Palmer, 1996). In other words, the goal of a test is to assess a test taker's language ability beyond the test context, and so if an L2 listening test

purports to assess listeners' ability to understand interactive, conversational spoken language (the TLU domain), then it should include spoken texts that involve interactive, conversational spoken language. The spoken texts used in the test should have similar phonological, linguistic, organizational, pragmatic and lexico-grammatical characteristics to real-world conversational language, and the pronunciation patterns in the test tasks should be similar to the pronunciation patterns of speakers in the TLU domain (Buck, 2001; Wagner, 2013b, 2014; Wagner & Toth, 2014). Using spoken texts in the test that have the same characteristics as spoken texts in the TLU domain should result in more valid inferences about the test takers' ability in that domain, while using spoken texts that utilize formal, over-enunciated spoken language can result in a narrow operationalization of the construct, which can present threats to the validity of the test (Messick, 1989, 1996).

In practice, however, many of the spoken texts used in L2 listening tests are very different from the spoken texts that test takers will encounter outside the test-taking context. In other words, the spoken texts used in the tests are not representative of the texts in the TLU domain. Wagner (2013b, 2014) describes how texts that are used in L2 listening tests are developed. Generally, test developers have test task specifications that dictate the number of texts to be used, their designated genre and length, and the number and types of response items to include for each. As a result, it is usually more efficient and practical for test developers to create a spoken text that corresponds to these specifications than to identify or record authentic texts. Thus, test developers usually create a scripted text (planned, written, revised, edited and polished) that is then read aloud by voice actors. The resulting scripted texts are often different from the unscripted, real-world spoken language of the TLU domain of interest.

Differences between scripted and unscripted texts

There has been extensive research on the characteristics of scripted and unscripted texts, and the results are summarized here (see Gilmore, 2007, for an analysis of how 'authentic' spoken texts differ from spoken texts commonly used in L2 materials). Wagner (2014), Wagner and Toth (2014) and Wagner and Wagner (2016) outline major categories of difference between texts that are planned, scripted and read aloud, and unplanned, spontaneous speech, where the speaker composes and utters the text virtually simultaneously. These include: articulatory/phonological characteristics (i.e. connected speech); organizational/planning characteristics; spoken grammar; oral lexicon; rate of speech; and hesitation phenomena. Among these, connected speech is the category most relevant for investigating pronunciation's role in L2 listening assessment, but the other five categories will also be briefly reviewed here as they are also secondarily related to the issue of pronunciation in L2 listening assessment.

Connected speech

The process of articulating rapid speech leads to phonological modifications that differ from citation forms in oral language. Brown and Kondo-Brown (2006) explain that articulated speech is affected by a number of processes, including word stress, sentence timing and stress, reduction, citation and weak forms of words, elision, intrusion, assimilation, juncture and contraction. These processes result in connected speech, which is typical of most real-world speaking events. It is also commonly accepted that the formality of a speech event affects pronunciation and connected speech: the more formal the event, the more careful and conscious the speaker is of enunciation, so that connected speech is less likely to occur. However, Brown and Kondo-Brown (2006) stress that, although connected speech is more common in informal contexts, it occurs in all registers and styles. Speakers can consciously attend to their pronunciation to reduce connected speech with a goal of clearer enunciation and greater intelligibility (Ito, 2006; Ladefoged & Johnson, 2010; Mora & Darcy, this volume). This is especially true when the speaker reads a text aloud, rather than composing and speaking a text simultaneously (Chafe, 1982; Haviland & Clark, 1974).

Organizational/planning characteristics

When scripted texts are read aloud, the speaker does not have to plan and speak simultaneously. Consequently, these texts often have a more formal, linear organization in the presentation of propositional content which reflects the planning and editing of the writer(s) of the text. In contrast, unscripted spoken texts tend to be less linearly organized because of the cognitive constraints involved in composing and uttering the text simultaneously (Chafe, 1982). As a result, unscripted spoken texts usually have many more digressions, false starts, redundancies and hesitation phenomena, including filled and unfilled pauses (Chafe, 1982; Haviland & Clark, 1974; Rost, 2011). This idea of hesitation phenomena is examined in more detail below.

Spoken grammar

Numerous researchers (e.g. Chafe, 1982, 1985; Halliday, 1985) have described how the grammatical characteristics of written (scripted) language differ from unscripted spoken language, including a greater use of complex syntactic structures like embedded clauses, agentless passives and nominalizations. In contrast, unscripted spoken texts generally have shorter idea units, more run-on sentences and less complex syntax. Indeed, corpus linguistic research has documented marked differences between the grammatical system of real-world spoken language and formal, scripted (written) language (Biber, 1988, 2006; Biber & Gray, 2013; McCarthy & Carter, 1995, 2001).

Oral lexicon

Similarly, Brown (1995) and Chafe (1985) have explained how written language generally contains less slang and colloquialisms than more

spontaneous, spoken language, in part because oral language tends to be less formal than written language. In addition, corpus linguists have shown that a speaker's oral and written lexicons can differ markedly from one another. McCarthy (2010) found that there was only about a 65% overlap between the 2000 most common words in a spoken versus a written corpus, and noted the importance of 'turn-openers' and 'turn-closers', like *yeah*, *oh*, and *mm*, in informal conversation. While these turn-openers and turn-closers can perform a number of functions, including as backchannels, and as turn-holding interactional strategies that the speaker uses while considering what to say and how to say it, they are much less common in scripted language, and most L2 learners (and teachers) would probably identify them as slang or colloquialisms.

Speech rate

Speech rate is generally defined as a measure of how quickly a person is speaking, and is often measured as the number of words, syllables or phonemes divided by the duration of the speech (Cucchiarini *et al.*, 2010). It is widely accepted that the rate of speech can affect L2 listening comprehension; the research has found that spoken texts delivered at a faster speech rate are more difficult for L2 listeners to comprehend than texts delivered at a slower rate (e.g. Griffiths, 1992; Kelch, 1985; Zhao, 1997). This is generally attributed to increased processing time, in that a slower speech rate allows the L2 listener more processing time. However, another consideration relevant to this study is the idea that a speaker's rate of speech can be influenced by his or her attempt to enunciate carefully. The fact that it is easier for a speaker to enunciate clearly when he or she speaks more slowly might make the text more intelligible (and comprehensible). This could partly explain why oral texts produced by highly proficient speakers of the target language and delivered at a faster rate are more difficult for L2 listeners to comprehend than texts delivered at a slower rate.

Hesitation phenomena

Hesitation phenomena are the filled and unfilled pauses, false starts, hesitations, redundancies and repeats that are characteristic of spontaneous spoken language. Because of the real-time nature of unplanned speech, where the speaker composes and utters speech almost simultaneously, hesitation phenomena can occur as the speaker searches for what to say and how to say it (Chafe, 1985; Wagner, 2014), or decides to rephrase an utterance (McCarthy, 2005). In spontaneous speech, filled pauses seem to be more common than unfilled pauses, while in scripted texts that are read aloud, unfilled pauses seem to be more common than filled pauses (Cucchiarini *et al.*, 2010; Wagner & Wagner, 2016).

How these hesitation phenomena affect L2 listeners' ability to understand spoken texts is a matter of debate. On the one hand, Griffiths (1991) explains that L2 learners might have difficulty in processing some types of hesitation

phenomena as they try to assign semantic meaning to filled pauses (*uh*, *you know*). Indeed, empirical studies such as Voss (1979) and Griffiths (1991) have found that L2 learners are less able to comprehend spoken texts with filled pauses than texts without them. Freedle and Kostin (1999) also examined the influence of hesitation phenomena on L2 listening performance, and found that texts with both filled pauses (e.g. *um* or *er*) and unfilled pauses of one second or more were actually more difficult for L2 test takers to comprehend than texts without such pauses. This was contrary to what they had hypothesized, and they concluded that 'apparently any disruption in the coherent reception of a speaker's ideas made it harder to process the message' (Freedle & Kostin, 1999: 18). However, while the findings of Voss (1979) and Griffiths (1991) focused on filled pauses, Freedle and Kostin grouped filled and unfilled pauses together because of their low frequency in the data, so the effects of a particular type of pause could not be ascertained. In addition, their pauses were purposefully inserted by the trained native English speakers of the texts to make them sound more authentic, so it is unclear whether these pauses were truly similar to and representative of the types of pauses found in real-world spoken language.

Alternatively, one could argue that hesitation phenomena might actually facilitate comprehension, in that the pauses and false starts would allow L2 listeners 'extra time to process what they hear' (Vandergrift & Goh, 2012: 154). This would possibly explain the findings reported earlier that a slower speech rate often leads to increased L2 comprehension. Indeed, Blau (1990) found that L2 learners scored higher on listening tests involving spoken texts with blank pauses mechanically inserted at normal discourse boundaries, than did learners hearing the same texts without these pauses. Similarly, Parker and Chaudron (1987) found that texts that included repetitions and redundancies (i.e. repeated phrases and clauses within the text) led to increased L2 listening comprehension. Again, this might be because these features effectively slowed down the speech rate. However, if the L2 listener is actively trying to decode the filled pauses and extract semantic information from them (as Voss, 1979, and Griffiths, 1991, found), or if the hesitation phenomena disrupt text processing, then the resulting slower speech rate might not actually benefit L2 listeners.

In summary, there is a broad consensus that the phonological/articulatory characteristics of unplanned spoken texts (i.e. connected speech that includes reduction, elision, intrusion, assimilation, juncture, etc.) make unscripted spoken texts more difficult for L2 listeners than the over-enunciated speech typical of scripted texts. Similarly, evidence suggests that spoken texts delivered at a faster rate are generally more difficult for L2 listeners than spoken texts delivered at a slower rate. However, even this issue is far from clear, because the hesitation phenomena typical of unplanned spoken texts (i.e. pauses and redundancies) can also serve to reduce the speech rate and, thus, potentially facilitate comprehension. The literature also suggests

that the organizational patterns, spoken grammar and oral lexicon characteristic of unscripted, spontaneous language can present difficulties for L2 listeners. Finally, spoken texts with the properties of unplanned spoken language might be more difficult for L2 listeners if they have not been exposed to or taught this kind of language. Indeed, the literature suggests that it is not necessarily the characteristics of unscripted spoken language that present difficulties for L2 listeners, but rather a lack of exposure to such texts, or instruction on and strategies for how to process them. Numerous researchers (e.g. Dupuy, 1999; Field, 2008; Gilmore, 2007; Meinardi, 2009; Wagner, 2014; Wagner & Toth, 2014) have argued that exposing L2 learners to unscripted oral texts and drawing learners' attention to their features so that learners will notice them in subsequent input is effective in developing the ability to comprehend unscripted, real-world spoken language. Yet for many L2 learners, especially foreign language learners, much of the spoken input they are exposed to comes from L2 textbook materials, and the types of spoken texts used in L2 textbooks have been found to differ extensively from real-world spoken language (Flowerdew & Miller, 1997; Gilmore, 2004, 2007; Thompson, 2003). Similarly, the spoken texts used in L2 listening tests seem to consist almost entirely of scripted texts that are read aloud (Wagner, 2013b), even though it is entirely feasible to use unscripted or semi-scripted spoken texts (Clark, 2014).

Learners' attitudes towards the use of authentic spoken texts

Field (2008) and Meinardi (2009) have argued that using authentic, unscripted texts for L2 listening instruction can enhance positive affect and motivation, which in theory should lead to more positive learning outcomes. This issue has received increasing attention in the field, and materials developers have seemed more concerned about the authenticity of their listening tasks in recent years. Gilmore (2004) examined the spoken texts used in ESL/EFL textbooks, and while most used inauthentic texts with few characteristics of real-world spoken language, more recent textbooks seem to be incorporating at least some natural discourse features in their texts. The issue has also begun receiving more attention in ESL/EFL teacher training materials (e.g. Brown, 2012). While many researchers have argued for authentic, unplanned spoken texts in L2 classrooms, some have expressed concern about the level of difficulty they entail, especially for lower proficiency learners. Guariento and Morley (2001) asserted that using such texts can lead to confused and frustrated learners and ultimately poor learning outcomes. Similarly, Richards (2006) cautioned about the 'myth of authenticity', arguing that authentic spoken texts for L2 instruction were difficult to obtain and even more difficult to implement without substantial modification.

A surprisingly limited number of empirical studies investigating language learners' attitudes towards unscripted texts in L2 materials have been

conducted, and the results are mixed, in part because the studies have used very different methodologies and examined learners of differing proficiency levels, but also perhaps because they have had relatively small participant numbers. Kmiecik and Barkhuizen (2006) surveyed 17 ESL learners, and found more negative attitudes towards authentic texts, because learners struggled with the speed of the input and the difficulty of the vocabulary. Peacock (1997) surveyed 31 EFL learners, who reported that even though they found authentic listening materials more motivating than artificial materials, they also found them less interesting. Gallien *et al.* (2000) surveyed 48 learners of French and German as a foreign language, who reported that simplified texts were more 'appealing' than authentic spoken texts.

Furthermore, one justification commonly made for authentic spoken L2 materials is that such material will be motivating for students. However, this is not fully supported by empirical data, and there seem to be no studies that investigate learners' attitudes towards authentic, unscripted spoken input in L2 listening tests. Perhaps more importantly, there is also very little empirical evidence examining the extent to which L2 learners are even aware of the differences between unscripted, real-world spoken language, and the scripted and polished spoken texts often found in L2 materials. Nonetheless, test takers' attitudes and beliefs about testing materials can have a real influence on their scores, as test taker affect can impact motivation and performance. In addition, the materials used in L2 tests (and the test takers' attitudes towards them) can contribute to a washback effect, both positively and negatively, on stakeholders including test takers, teachers and educational systems (Buck, 2001; Wagner, 2014; Wagner & Wagner, 2016). If a high-stakes test uses unscripted spoken texts for L2 listening assessment, then it is more likely that curriculum planners, materials developers and classroom teachers will likewise use unscripted spoken texts in their materials for L2 learners.

The Current Study

The current study explores test takers' awareness and beliefs about the types of spoken texts used in an L2 listening test. It is part of a larger investigation of unscripted spoken texts in L2 listening comprehension assessment. As reported in Wagner and Toth (2014), two comparable groups of L2 Spanish learners took an L2 listening test. For one group, the spoken texts were unscripted, and consequently had many of the organizational, phonological and fluency characteristics found in spontaneous, real-world language, as well as extensive instances of connected speech and hesitation phenomena. The second group took the same L2 listening test, except that the spoken texts were scripted and lacked most of the characteristics of unplanned spoken language. As hypothesized, the group of 86 test takers that listened to the scripted texts scored 8.4% higher on the listening

comprehension test than the group of 85 test takers who listened to the unscripted texts, and this difference was statistically significant (Wagner & Toth, 2014). However, we were not only interested in how the two different groups would perform on the test, but we also wanted to examine the extent to which the test takers were aware of the organization, phonology and fluency characteristics of the different spoken texts. Thus, after completing the test, test takers in both groups were surveyed through the use of a written questionnaire about the spoken texts used in the test. The following research question was addressed: What are the test takers' beliefs about the characteristics of the spoken language used on an L2 listening comprehension test?

Methodology

For the questionnaire data, the independent variable was the type of audio-text used in the listening comprehension test: 'unscripted' versus 'scripted'. The dependent variables were group scores on five different sub-scales of the questionnaire that asked the participants about their views of the spoken texts used in the listening test. A series of independent sample *t*-tests assessed how the independent variable affected the group scores on the sub-scales of the questionnaire.

Participants

This study involved 171 learners of Spanish as a foreign language (SFL) at a large American public university. All were students in an intermediate-level, fourth-semester Spanish course entitled 'Conversational Review', which focused on speaking and listening skills. There were 14 classes of 'Conversational Review' in the study, which were randomly assigned to one of two groups: seven classes were assigned to the unscripted group, and seven to the scripted group. Of the 85 test takers in the unscripted group, 81 listed English as their L1, and four listed a language other than English. None had Spanish as their L1, although two listed a Romance language (Romanian). For the unscripted group, the average age was 20.24 years and 59% were female. For the 86 participants in the scripted group, 78 listed English as their L1, and seven listed a language other than English. None had Spanish or a Romance language as their L1. The average age of the group was 20.45 years and 70% were female. A self-assessment was used to examine if the two groups had comparable perceptions of their L2 Spanish proficiency. The test takers used a six-point scale (1 = lower beginner, 2 = upper beginner, 3 = lower intermediate, 4 = upper intermediate, 5 = lower advanced, 6 = upper advanced). The two groups' self-assessments were very similar: 3.49/6.00 for the unscripted group and 3.53/6.00 for the scripted group.

Spoken texts

Two spoken texts were created for the study described in Wagner and Toth (2014). Two female L1 speakers of Peruvian Spanish were used; they were given the basic outlines for performing a role-play to create the two texts: one called 'A Room for Rent', and another called 'A Friend Goes on Vacation'. For the former, one speaker was a university student seeking to rent a room from the other; for the latter, one speaker gave instructions to the other for taking care of her house while she went on vacation. After reading the role-play instructions and considering what they might say for a few moments, the two speakers were instructed to speak as naturally as possible for approximately three to four minutes. The speakers then recorded the two texts.

After the unscripted texts were completed, the researchers transcribed them, and then revised and edited the transcripts to remove the pauses, false starts, hesitations, redundancies, overlaps and backchannels. This resulted in fewer speaker turns in the texts, and a more linear organizational scheme. Using these edited and polished transcripts, the same two native Spanish speakers were then instructed to read the transcripts aloud, and to simulate the types of spoken texts found in L2 classroom materials. They were instructed to be conscious of enunciating clearly, to avoid connected speech and overlapping with the other speaker, and not to speak too rapidly.

The resulting two versions of the spoken texts were thus equivalent in topic, content and information, and were spoken by the same speakers. They differed, however, in the presence or absence of connected speech and hesitation phenomena and their related organizational characteristics, as virtually all instances of overlapping talk, filled pauses, repeated phrases, backchannels and exclamatives from the unscripted texts were absent in the scripted versions (see Wagner & Toth, 2014). Thus, in the scripted text, related propositions spread over two or more turns (with interruptions from the other speaker) were consolidated into single turns, and all false starts, repetitions, backchannels, exclamatives and filled pauses were simply deleted. It should be noted, however, that the vocabulary and propositional content in the two texts was nearly identical. Furthermore, there was no slang or colloquial language used in either version of the texts, apart from the fillers and backchannels in the unscripted text.

Instruments

After the spoken texts were created, eight multiple-choice listening comprehension items were developed for each of the texts, resulting in a 16-item test. A 21-item questionnaire was administered to examine the test takers' beliefs and opinions about the spoken input they heard on the exam. It was developed and validated based on Wagner's (2010, 2013a) suggestions for

applied linguistics survey research. The questionnaire used Likert items with five choices: 5 = 'strongly agree', 4 = 'agree', 3 = 'no opinion', 2 = 'disagree' and 1 = 'strongly disagree'. The five sub-scales of the questionnaire (described below) were based on a review of literature on the use of unscripted spoken texts in L2 teaching and testing (e.g. Gilmore, 2007; Wagner, 2013b).

The first sub-scale, 'Authentic versus Modified', was a five-item, holistic measure of beliefs about whether the recordings used in the test were real-world spoken texts or scripted texts created specifically for L2 learners. For example, one item asked: 'The texts required me to listen to authentic spoken language, the same type of spoken language that is found in real life.' A group mean above 3 on this sub-scale indicates that test takers thought the spoken texts were authentic and unscripted, while a group mean below 3 indicates that they thought the texts were scripted and modified, and created for L2 learners.

The other four sub-scales, each composed of four items, asked about specific characteristics of the text. The second sub-scale, 'Pronunciation', asked test takers about how the speakers enunciated their speech. For example, one item asked: 'It was hard to understand the speakers because they did **not** enunciate well and did **not** speak clearly.' (This item, and a number of other negatively worded items were reverse-coded in the analysis.) Other options in this sub-scale included: 'The speakers spoke clearly and used very clear pronunciation, which made it easier to understand them'; 'The speakers pronounced each word clearly and distinctly'; and 'The speakers' pronunciation in the texts was similar to native Spanish speakers' pronunciation in real-life conversations.' A group mean above 3 on this sub-scale indicates that test takers thought the speakers enunciated normally and that their pronunciation was similar to real-world spoken language, while a group mean below 3 indicates that they thought the speakers over-enunciated and spoke extra clearly so that L2 listeners could understand them.

The third sub-scale, 'Speech Rate', asked about how quickly the speakers spoke in the texts. For example, one item asked: 'The speakers on the spoken texts spoke quickly, the same rate that native speakers normally use with each other.' A group mean above 3 on this sub-scale indicates that the test takers thought the speakers spoke quickly, similarly to highly proficient speakers in real-world contexts, while a group mean below 3 indicates that they thought the speakers spoke artificially slowly and deliberately.

The fourth sub-scale, 'Pauses and False Starts', asked test takers about the extent to which the speakers in the texts had hesitation phenomena in their speech. For example, one item asked: 'The speakers often had a lot of pauses, fillers (things like *'eh …'*, *'em …'*, *'este …'*, *'tú sabes …'*), and false starts in their speech.' A group mean above 3 on this sub-scale indicates that the test takers thought the spoken texts had pauses, fillers and false starts like those in real-life unplanned spoken communication, while a group mean below 3 indicates that they thought the spoken texts were rehearsed and read aloud and did not include hesitation phenomena found in unscripted language.

The fifth sub-scale, 'Use of Slang', asked whether test takers thought the speakers used slang in their speech. For example, one item asked: 'The speakers used slang and informal expressions that are found in real-life language.' As stated above, the vocabulary used in the two texts was virtually identical, and thus there was no difference in the amount of slang or colloquial language used in the scripts. Nevertheless, we decided to include this sub-scale on the questionnaire in order to examine the extent to which the learners associated unscripted language with colloquial or non-standard speech. This seemed relevant, given our experience with language learners referring to any non-standard speech as *slang*, with a somewhat negative connotation. A group mean above 3 on this sub-scale indicates that the test takers thought the speakers did use slang and colloquial language, while a group mean below 3 indicates that they thought the speakers did not use slang.

The initial 21-item questionnaire was created, and then piloted with a group of 14 learners in a 'Conversational Review' class. After completing the questionnaire, the 14 test takers were surveyed about the questionnaire items, and asked about any items they found particularly difficult or problematic. In addition, a statistical analysis of the responses was conducted. Based on these qualitative and quantitative analyses, a number of items were revised until the questionnaire resulted in its final form. A complete list of the questionnaire items is provided in the Appendix to this chapter.

Procedures

The researchers went to the 14 different 'Conversational Review' classes to administer the test and post-test questionnaire. Because these were low-proficiency learners, the directions for the listening comprehension test were given in English both on the audio-recording and in the test booklet. The test items were written in Spanish. The test took about 20 minutes to complete, after which the test takers completed the 21-item questionnaire. The questionnaire items were ordered randomly and were written in English. Test takers circled numbers corresponding with an answer of 'strongly disagree', 'disagree', 'no opinion/don't know', 'agree', and 'strongly agree'. The questionnaire took about 10 minutes.

Test takers were not told before they took the test or questionnaire what the purpose of the study was. Rather, they were told that the researchers were examining how L2 learners perform on a listening test. They did not know that there were two versions of the spoken texts used in the test, and thus when they completed the questionnaire they responded based only on the version that they had just heard.

Analyses

The internal consistency reliability for each of the five sub-scales on the questionnaire was estimated separately for each group using Cronbach's

alpha. The item-total correlation for each item with its overall sub-scale was also examined to see how reliably each item performed. Descriptive statistics were computed to examine the central tendency and dispersion of the two groups on each of the sub-scales. A series of independent sample *t*-tests was then conducted to see if the two groups' beliefs and impressions of the two texts differed; that is, the means for the two groups' scores on the five sub-scales (authentic versus modified, pronunciation, speech rate, pauses and false starts, and use of slang) were compared to see if the groups' beliefs about the five variables differed.

Results

Beliefs about the spoken texts

While 20 of the 21 items on the questionnaire performed well statistically, the item-total correlation for item 21 (part of the 'pronunciation' sub-scale) was very low for both groups. Test takers were asked: 'The speakers' pronunciation in the texts was similar to native Spanish speakers' pronunciation in real-life conversations.' In reviewing this item, it became apparent that it differed from the other pronunciation items in that it asked whether the speakers' pronunciation was similar to that of native speakers, while the other four items focused on clarity and enunciation. Because the native speaker item did not seem to be reliably measuring the same construct, it was deleted from the rest of the analysis.

For both the unscripted and scripted groups, each of the five sub-scales had a moderately high internal consistency. For the unscripted group, the reliability coefficient for each sub-scale was: authentic versus modified, $\alpha = 0.74$; pronunciation, $\alpha = 0.81$; speech rate, $\alpha = 0.80$; pauses and false starts, $\alpha = 0.68$; and use of slang, $\alpha = 0.80$. For the scripted group, the reliability for each sub-scale was: authentic versus modified, $\alpha = 0.80$; pronunciation, $\alpha = 0.73$; speech rate, $\alpha = 0.80$; pauses and false starts, $\alpha = 0.54$; and use of slang, $\alpha = 0.82$. While these reliability figures are relatively high, the coefficient for the pauses and false starts sub-scale for the scripted group is markedly lower.

The descriptive statistics for both groups' scores on the five sub-scales of the questionnaire were also calculated. As shown in Table 5.1, the mean scores on the five sub-scales are consistently higher for the unscripted group than the mean scores for the scripted input group. To reiterate, the means for each sub-scale are based on five-point scales with 3 as the mid-point, so a higher mean on these sub-scales indicates that the test takers thought the texts were more authentic, that the pronunciation was more like real life, that there was a more natural speech rate, that there were more pauses and false starts and, finally, that there was more slang.

Table 5.1 Descriptive statistics for the scripted and unscripted groups on the five sub-scales of the questionnaire ($n = 171$)

Variable	*Authentic vs. modified*		*Pronunciation*		*Speech rate*		*Pauses and false starts*		*Use of slang*	
	U	*S*	*U*	*S*	*U*	*S*	*U*	*S*	*U*	*S*
Mean rating	3.52	2.99	3.22	2.14	3.85	2.66	3.49	2.40	2.91	2.39
SD	0.55	0.67	0.86	0.63	0.69	0.81	0.64	0.49	0.62	0.60
Kurtosis	0.04	–0.81	–1.20	0.95	0.86	0.86	–0.49	0.62	–0.77	–0.72
Skewness	–0.42	–0.16	–0.26	0.91	–1.10	0.37	–0.11	–0.12	–0.13	–0.07
Reliability	0.74	0.80	0.81	0.73	0.80	0.80	0.68	0.54	0.80	0.82

Notes: U = unscripted group; S = scripted group. Mean ratings are based on a five-point scale. Reliability is calculated using Cronbach's alpha.

Between-group comparisons

In order to compare the two groups' means on the five sub-scales of the questionnaire, five independent-sample *t*-tests were conducted. Because using five *t*-tests raises the possibility of finding group differences when in fact there are none, a Bonferroni adjustment set the significant level for multiple comparisons at 0.01. The skewness and kurtosis figures given in Table 5.1 suggest that the data for both groups are normally distributed. Levene's test of homogeneity of variances was conducted on the five sub-scales and it was found that for only one of the sub-scales, the use of slang, could the variances be considered homogeneous. Therefore, on the other four sub-scales, the numbers reported in the *t*-tests will be for 'equal variances not assumed'. The two-tailed *t*-tests for all five comparisons of the two groups' scores on the questionnaire sub-scales were statistically significant: authentic versus modified, $t(168.64) = 5.70$, $p < 0.001$, $d = 0.86$; pronunciation, $t(154.04) = 10.45$, $p < 0.001$, $d = 1.43$; speech rate, $t(165.36) = 10.45$, $p < 0.001$, $d = 1.58$; pauses and false starts, $t(156.59) = 12.50$, $p < 0.001$, $d = 1.91$; and use of slang, $t(169) = 5.56$, $p < 0.001$, $d = 0.85$. As shown by Cohen's *d*-effect size values, the effect sizes for the five sub-scales were all large. These tests indicated that there was a statistically significant difference in the two groups' beliefs about the spoken input on the tests for each of the five sub-scales of the questionnaire.

Discussion

Our research question asked: 'What are the test takers' beliefs about the characteristics of the spoken language used on an L2 listening comprehension

test?' The results indicate that the two groups of test takers had very different beliefs about the texts based on whether they heard the unscripted or the scripted texts. Again, test takers were not informed about the purpose of the study before they took the test or questionnaire, and they did not know which type of text they had heard. Yet the two groups' responses to the questionnaire differed significantly on all five sub-scales.

The unscripted group's score was more than a half-point higher (3.52 versus 2.99) than the scripted group on the first sub-scale, which asked whether participants thought the texts they heard were authentic, natural and representative of real-world spoken language. This statistically significant result indicates that learners could indeed distinguish authentic spoken texts from those created especially for L2 learners.

The second sub-scale focused on the extent to which test takers thought the texts had the pronunciation patterns and characteristics found in real-life Spanish conversations (i.e. if the speakers enunciated on the texts as they would in real conversation). The unscripted group's score of 3.22 was more than a full point higher than the scripted group's score of 2.14, which was the lowest of any score on the five sub-scales. Thus, participants in the scripted group were well aware that the pronunciation they heard was different from real-world language, and that the speakers were enunciating more clearly than they would in a real-world context. Similarly, the unscripted group's score on the speech rate sub-scale was more than a full point higher than the scripted group (3.85 and 2.66, respectively), which yielded the largest difference for any of the five sub-scales. This meant that the unscripted group participants agreed with statements affirming that the speakers on the texts spoke quickly, as native speakers do when conversing. For the fourth sub-scale, which asked whether the spoken texts had pauses and false starts similar to those of real-world language, the unscripted group again scored more than a point higher (3.49) than the scripted group (2.40), suggesting that listeners perceived the hesitation phenomena that were present in the unscripted texts but virtually absent in the scripted texts. However, the reliability coefficient was much lower for the pauses and false starts sub-scale for the unscripted and scripted groups ($\alpha = 0.68$ and $\alpha = 0.54$, respectively), so the results must be interpreted with caution.

The results of the scores on the fifth and final sub-scale, 'use of slang', are difficult to interpret. This sub-scale asked participants about how much slang and colloquial language the speakers used in the text. The mean score of 2.91 was the lowest of the five sub-scales for the unscripted group and below, in fact, the mid-point of the sub-scale. Similarly, the mean score of 2.39 for the scripted group was the lowest of the five sub-scales. While this was the smallest difference in means for any of the five sub-scales, it was still statistically significant. These scores are difficult to explain because the vocabulary in the two texts was virtually identical, with no lexical

modifications made to the scripted text apart from the removal of filled pauses. One possible reason for the difference in scores is that, because the unscripted group perceived their text to be more natural, they might have assumed the speakers were using slang and colloquialisms (including the fillers such as *'um'*, *'este'* and *'o sea'*). Similarly, because the scripted group perceived the texts as being unnatural with overly formal enunciation, they might have assumed that the speakers would be less likely to use slang and colloquial language.

A limitation of this study is that while it compared two groups' beliefs about the texts they heard, each group heard only one type of text. A counterbalanced design in which each group heard and rated both types of texts would have been stronger. Nevertheless, our findings suggest that learners can distinguish spoken texts made especially for L2 learners from unplanned, unscripted speech that reflects real-world spoken language. There does not seem to be any literature that has specifically focused on L2 learners' ability to detect if a spoken text is scripted or unscripted, so these results must be seen as exploratory. As reported in Wagner and Toth (2014), the learners in the unscripted group scored lower on the comprehension test than the unscripted group. It is not surprising, then, that these learners would report that the texts seemed similar to the authentic spoken language of native speakers. Because of the learners' relative difficulty in comprehending and processing the text, we can speculate that they associated it with authentic, unscripted speech, and thus perceived the speakers as talking quickly, using slang and colloquialisms, not enunciating clearly, and employing numerous pauses and fillers. This would mirror the results of Kmiecik and Barkhuizen (2006), who found that ESL learners had more negative attitudes towards authentic spoken texts due to comprehension difficulties arising from a high speech rate and the use of unfamiliar vocabulary. Likewise, our results reflect Gallien *et al.*'s (2000) study, where FL learners found simplified texts more appealing than authentic texts, in part due to ease of comprehension.

It seems unlikely that the listeners in this study were conscious of many of the organization, phonology and fluency characteristics of the spoken texts while they were listening to them. Rather, at least for some participants, the items in the questionnaire likely forced them to think about the different characteristics, and test takers who had difficulty with the texts may have equated the challenge they faced with a particular text type. This would perhaps explain the anomaly of the 'use of slang' sub-scale. Again, the test takers in the unscripted group rated their texts as having more slang and colloquial language than test takers in the scripted group, even though the vocabulary was virtually identical in both text versions.

In order to examine this hypothesis, we carried out a post hoc analysis, in which we divided the two test taker groups into 'high-comprehenders', who scored above the median on the listening comprehension test, and

'low-comprehenders', who scored at or below the median. For the scripted group, there was no difference among high- and low-comprehenders' mean scores on any of the five sub-scales. For the unscripted group, however, the low-comprehenders' mean score (3.02, $SD = 0.56$) on the 'use of slang' sub-scale was significantly higher than the high-comprehenders' mean score (2.71, $SD = 0.71$), $t(83) = 2.31$, $p = 0.024$. In addition, the low-comprehenders' mean score (4.02, $SD = 0.56$) was significantly higher than the high-comprehenders' mean score (3.56, $SD = 0.79$) on the 'speech rate' sub-scale, $t(83) = 3.178$, $p = 0.002$. There was no difference between the high- and low-comprehenders' scores on the three other sub-scales. Thus, these results suggest that the test takers with lower comprehension scores in the unscripted group might have perceived the unscripted text as having slang and colloquial language because there was a good amount of vocabulary they could not decipher. Similarly, the low-comprehenders might have attributed their inability to comprehend the spoken texts to the seemingly rapid speech rate. These conclusions are merely speculative, however. Because there is so little research on the extent to which L2 listeners can perceive the organization, phonology, and fluency characteristics of unscripted spoken texts, more work is obviously needed in this area.

Implications and Conclusion

It is almost universally acknowledged that the goal for adult L2 learners in regard to their own pronunciation is intelligibility (e.g. Ballard & Winke, this volume; Harding, this volume; Isaacs, 2013; Isaacs & Trofimovich, 2012; Trofimovich & Isaacs, this volume), given that a fully 'nativelike' pronunciation is usually an unrealistic, inappropriate expectation for adult learners. This belief is confirmed by the fact that 'sounds like a native speaker' is no longer used as a descriptor on pronunciation rubrics/ratings scales. Similarly, it is almost universally acknowledged that real-world spoken language contains connected speech and hesitation phenomena that are not the result of 'lazy' or 'sloppy' pronunciation, but are in fact a normal, necessary and appropriate result of articulating spontaneous spoken language. And yet, believing that they are making listening comprehension more accessible to learners by maximizing intelligibility, L2 materials and test developers continue using unrealistic and inauthentic models of pronunciation in the spoken texts in their materials. Clark (2014) has demonstrated that it is feasible to commission semi-scripted spoken texts for L2 listening tasks, yet the vast majority of L2 listening tests use spoken texts with pronunciation involving formal, over-enunciated citation forms of language that differ dramatically from spontaneous, real-world spoken language. Indeed, by perpetuating inauthentic speech models at the expense of appropriate models of real-world speech, we believe that learners are disadvantaged in that they

acquire inaccurate perceptions of what L2 speakers should sound like and consequently feel unprepared to engage in discourse beyond the classroom. As in Wagner (2014) and Wagner and Toth (2014), we suggest that simplification strategies other than text modification be used to make real-world L2 speech accessible and intelligible to learners, including the careful management of: (a) text length; (b) the targets of attentional focus; (c) the intended depth of learners' comprehension; (d) the number of listening rounds; and (e) opportunities for hypothesizing, feedback and knowledge consolidation. Ultimately, teachers must help learners cope with their inability to understand *everything*, so that learners can build confidence in their ability to understand *something* and thus establish a grounding in comprehension that will sustain them in natural conversation. Similarly, drawing learners' attention to the characteristics of unplanned spoken language should also help learners notice and attend to these characteristics in subsequent spoken input, both inside and outside the language classroom, which corresponds very closely with Vandergrift and Goh's (2012) metacognitive approach to L2 listening instruction.

This study has demonstrated that L2 learners can identify spoken texts that are specially created for L2 learners, and even distinguish the organization, phonology and fluency characteristics of these texts from spontaneous, real-world spoken language. L2 test developers must therefore consider aspects of pronunciation not only when developing speaking tests, but also when developing L2 listening tests. They should include unscripted spoken texts in L2 listening tests, because doing so will result in better domain coverage (i.e. texts that are more reflective of the spoken texts in the real world) and more valid inferences about test takers' ability to understand real-world spoken language. In addition, the inclusion of these types of spoken texts can have a positive washback effect not only on test takers, but also for the larger educational systems that prepare students to take them, by promoting the use of unscripted, spontaneous texts in the L2 classroom and materials. If students know that these types of texts will appear in L2 listening tests, then they should be more receptive to their use in the classroom, even if they perceive them as initially more difficult. Similarly, L2 teachers and curriculum and materials developers should regularly implement unscripted spoken texts in L2 listening tasks, especially with more advanced learners, but even beginning learners can benefit from being exposed to these types of texts. As our results suggest, learners can readily tell when they are hearing inauthentic speech. If indeed the possibility of engaging with real-world spoken language strengthens learner motivation, as Peacock (1997) suggests, while also provoking anxiety, then our primary instructional concern should be providing sufficient support during experiences of real-world texts to make comprehension possible and thereby build among learners a repertoire of successful experiences that ultimately leads to a noticing of and familiarity with unscripted, spontaneous communication.

References

Bachman, L. and Palmer, A. (1996) *Language Testing in Practice*. Oxford: Oxford University Press.

Biber, D. (1988) *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (2006) *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

Biber, D. and Gray, B. (2013) Discourse characteristics of writing and speaking tasks types on the TOEFL iBT Test: A lexico-grammatical analysis. TOEFL Research Report No. RR-13-04. Princeton, NJ: Educational Testing Service.

Blau, E. (1990) The effect of syntax, speed, and pauses on listening comprehension. *TESOL Quarterly* 24, 746–753.

Brown, G. (1995) *Speakers, Listeners, and Communication*. Cambridge: Cambridge University Press.

Brown, J.D. (ed.) (2012) *New Ways in Teaching Connected Speech*. Alexandria, VA: TESOL International Association.

Brown, J.D. and Kondo-Brown, K. (2006) Introducing connected speech. In J.D. Brown and K. Kondo-Brown (eds) *Perspectives on Teaching Connected Speech to Second Language Speakers* (pp. 1–16). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

Buck, G. (2001) *Assessing Listening*. Cambridge: Cambridge University Press.

Chafe, W. (1982) Integration and involvement in speaking, writing, and oral literature. In D. Tannen (ed.) *Spoken and Written Language: Exploring Orality and Literacy* (pp. 35–53). Norwood NJ: Ablex.

Chafe, W. (1985) Linguistic differences produced by differences between speaking and writing. In D. Olson, D. Torrance and A. Hildyard (eds) *Literacy Language and Learning* (pp. 105–123). Cambridge: Cambridge University Press.

Clark, M. (2014) The use of semi-scripted speech in a listening placement test for university students. *Papers in Language Testing and Assessment* 3 (2), 1–26.

Cucchiarini, C., van Doremalen, J. and Strik, H. (2010) Fluency in non-native read and spontaneous speech. *Proceedings of Disfluency in Spontaneous Speech (DiSS) and Linguistic Patterns in Spontaneous Speech (LPSS) Joint Workshop* (pp. 15–18).

Dupuy, B. (1999) Narrow listening: An alternative way to develop and enhance listening comprehension in students of French as a foreign language. *System* 27, 351–361.

Field, J. (2008) *Listening in the Language Classroom*. Cambridge: Cambridge University Press.

Flowerdew, J. and Miller, L. (1997) The teaching of academic listening comprehension and the question of authenticity. *English for Specific Purposes* 16, 27–46.

Freedle, R. and Kostin, I. (1999) Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing* 16, 2–32.

Gallien, C., Hotho, S. and Staines, H. (2000) The impact of input modifications on listening comprehension: A study of learner perceptions. *JALT Journal* 22, 271–295.

Gilmore, A. (2004) A comparison of textbook and authentic interactions. *ELT Journal* 58, 363–374.

Gilmore, A. (2007) Authentic materials and authenticity in foreign language learning. *Language Teaching* 40, 97–118.

Griffiths, R. (1991) The paradox of comprehensible input: Hesitation phenomena in L2 teacher talk. *JALT Journal* 13, 23–41.

Griffiths, R. (1992) Speech rate and listening comprehension: Further evidence of the relationship. *TESOL Quarterly* 26, 385–391.

Guariento, W. and Morley, J. (2001) Text and task authenticity in the EFL classroom. *ELT Journal* 55, 347–353.

Halliday, M.A.K. (1985) *Introduction to Functional Grammar*. London: Edward Arnold.
Haviland, S. and Clark, H. (1974) What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior* 13, 512–521.
Isaacs, T. (2013) Pronunciation. In *Cambridge English Centenary Symposium on Speaking Assessment* (pp. 13–15). Cambridge: Cambridge English Language Assessment.
Isaacs, T. (2014) Assessing pronunciation. In A. Kunnan (ed.) *Companion to Language Assessment* (Vol. 1, pp. 140–155). Oxford: Wiley-Blackwell.
Isaacs, T. and Trofimovich, P. (2012) Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition* 34, 475–505.
Ito, Y. (2006) The significance of reduced forms in L2 pedagogy. In J.D. Brown and K. Kondo-Brown (eds) *Perspectives on Teaching Connected Speech to Second Language Speakers* (pp. 17–26). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.
Kelch, K. (1985) Modified input as an aid to comprehension. *Studies in Second Language Acquisition* 7 (1), 81–90.
Kmiecik, K. and Barkhuizen, G. (2006) Learner attitudes towards authentic and specially prepared listening materials: A mixed message? *TESOLANZ Journal* 14, 1–15.
Ladefoged, P. and Johnson, K. (2010) *A Course in Phonetics* (6th edn). Boston, MD: Cengage.
McCarthy, M. (2005) Fluency and confluence: What fluent speakers do. *The Language Teacher* 29, 26–28.
McCarthy, M. (2010) Spoken fluency revisited. *English Profile Journal* 1, 1–15.
McCarthy, M. and Carter, R. (1995) Spoken grammar: What is it and how can we teach it? *ELT Journal* 49, 207–218.
McCarthy, M. and Carter, R. (2001) Ten criteria for a spoken grammar. In E. Hinkel and S. Fotos (eds) *New Perspectives on Grammar Teaching in Second Language Classrooms* (pp. 51–75). Mahwah, NJ: Lawrence Erlbaum.
Meinardi, M. (2009) Speed bumps for authentic listening material. *ReCALL* 21, 302–318.
Messick, S. (1989) Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18 (2), 5–11.
Messick, S. (1996) Validity and washback in language testing. *ETS Research Report Series* 1, 1–18.
Parker, K. and Chaudron, C. (1987) The effects of linguistic simplification and elaborative modifications on L2 comprehension. *University of Hawai'i Working Papers in ESL* 6, 107–133.
Peacock, M. (1997) The effect of authentic materials on the motivation of EFL learners. *ELT Journal* 51, 144–156.
Richards, J. (2006) Materials development and research: Making the connection. *RELC Journal* 37, 5–26.
Rost, M. (2011) *Teaching and Researching Listening* (2nd edn). Harlow: Pearson Education.
Thompson, S. (2003) Text-structuring metadiscourse, intonation, and the signaling of organization in academic lectures. *Journal of English for Academic Purposes* 2, 5–20.
Vandergrift, L. and Goh, C. (2012) *Teaching and Learning Second Language Listening: Metacognition in Action*. New York: Routledge.
Voss, B. (1979) Hesitation phenomena as sources of perceptual errors for non-native speakers. *Language and Speech* 22, 129–144.
Wagner, E. (2010) Survey research in applied linguistics. In B. Paltridge and A. Phakiti (eds) *Continuum Companion to Second Language Research Methods* (pp. 22–38). London: Continuum.
Wagner, E. (2013a) Surveys. In C. Chapelle (ed.) *The Encyclopedia of Applied Linguistics* (pp. 5470–5475). Oxford: Wiley-Blackwell.
Wagner, E. (2013b) Assessing listening. In A. Kunan (ed.) *Companion to Language Assessment*. Oxford: Wiley-Blackwell.

Wagner, E. (2014) Using unscripted spoken texts in the teaching of second language listening. *TESOL Journal* 5, 288–311.
Wagner, E. and Toth, P. (2014) Teaching and testing L2 Spanish listening using scripted versus unscripted texts. *Foreign Language Annals* 47, 404–422.
Wagner, E. and Wagner, S. (2016) Scripted and unscripted spoken texts used in listening tasks on high stakes tests in China, Japan, and Taiwan. In V. Aryadoust and J. Fox (eds) *Current Trends in Language Testing in the Pacific Rim and the Middle East: Policies, Analyses, and Diagnoses* (pp. 103–123). Newcastle upon Tyne: Cambridge Scholars Publishing.
Zhao, Y. (1997) The effects of listeners' control of speech rate on second language comprehension. *Applied Linguistics* 18, 49–68.

Appendix: Post-test Questionnaire

Authentic versus modified sub-scale

(1) The spoken texts were **not** authentic; they were specially created for students learning Spanish.
(2) The texts used authentic spoken input, like is found in real life.
(3) The speakers planned and practised what they were going to say and read from transcripts.
(4) The texts required me to listen to authentic spoken language, the same type of spoken language that is found in real life.
(5) The texts that were used did **not** have authentic spoken input.

Pronunciation sub-scale

(1) The speakers spoke clearly and used very clear pronunciation, which made it easier to understand them.
(2) The speakers pronounced each word clearly and distinctly.
(3) It was hard to understand the speakers because they did **not** enunciate well and did **not** speak clearly.
(4) The speakers' pronunciation in the texts was similar to native Spanish speakers' pronunciation in real-life conversations (excluded from analyses due to low item-total correlation of this item with the other items in the pronunciation sub-scale).

Speech rate sub-scale

(1) The speakers on the spoken texts spoke quickly, the same rate that native speakers normally use with each other.
(2) The speakers in the spoken texts spoke slowly and enunciated each word.
(3) The speakers spoke slowly and clearly so that the listeners would be able to understand them.
(4) The speakers spoke quickly, at the same rate as native speakers in real-life conversations.

Use of slang sub-scale

(1) The speakers used only formal, standard Spanish, with **no** slang.
(2) The spoken texts often had slang and colloquial speech in them.
(3) The speakers did **not** use slang or informal expressions when they were speaking.
(4) The speakers used slang and informal expressions that are found in real-life language.

Pauses and false starts sub-scale

(1) The speakers often had a lot of pauses, fillers (things like 'eh …', 'em …', 'este …', 'tú sabes …'), and false starts in their speech.
(2) Because the speakers planned what they were going to say and read from a transcript, there were few pauses, false starts, and fillers (things like 'eh …', 'em …', 'este …', 'tú sabes …') in the spoken texts.
(3) I could tell that the speakers were reading from a transcript, because they did **not** have any pauses, false starts or mistakes in their speech.
(4) There were pauses, fillers and false starts in the texts, suggesting that the speakers did **not** plan what they were going to say, and were **not** reading from transcripts.

Part 3

Perspectives on Pronunciation Assessment From Psycholinguistics and Speech Sciences

6 The Relationship Between Cognitive Control and Pronunciation in a Second Language

Joan C. Mora and Isabelle Darcy

Introduction

In today's world of increasing mobility, millions of people are nonnative speakers of languages they use daily. Efficient oral communication skills in a second language (L2) have become crucial. Yet the vast majority of us who attempt to learn a new language will be likely to have a pronunciation that sounds foreign accented. A foreign accent can sometimes hinder communication and lower intelligibility and is associated with many unconscious negative stereotypes (e.g. Gluszek & Dovidio, 2010). Concurrently, a long-standing challenge to adults in foreign language education is pronunciation instruction. Two reasons may account for this: first, intelligible pronunciation is difficult to *learn* for most adults and, secondly, intelligible pronunciation is difficult to *teach*, and for many learners, pronunciation instruction appears inefficient. Consequently, many second/foreign language programmes sideline pronunciation in their curricula and many teachers are ill-prepared to teach it (Darcy *et al.*, 2012). In addition, learners display variable outcomes in developing intelligible L2 pronunciation as a result of instruction. Everything else being equal, some show little progress even over long time periods, whereas others appear to 'get it' much sooner, and clearly benefit from pronunciation instruction (Thomson & Derwing, 2015). Generally speaking, pronunciation is the area of language with the largest individual variation in performance, compared to grammar or vocabulary, with some learners speaking the L2 with a strong foreign accent even after years of L2 immersion and only very few advanced learners being able to sound nativelike (Abrahamsson & Hyltenstam, 2009; Moyer, 1999). The reasons for this heterogeneity are still largely unknown. Of course, certain conditions of

learning (such as the native language (L1) background, or age and length of learning) are known to affect pronunciation 'success'. However, individual differences often remain after these variables are controlled (Bradlow *et al.*, 1997; Golestani & Zatorre, 2009). The investigation of individual differences in language has generally used global measures of success (e.g. overall test scores), while few studies have scrutinized the phonological skills of learners (e.g. Darcy *et al.*, 2015; Hu *et al.*, 2013). There is, therefore, a key gap in our knowledge base with respect to which factors underlie individual differences in L2 phonological development.

Although there is evidence showing that successful phonological acquisition is linked to better cognitive control, which refers to the general-purpose control mechanisms that regulate information processing skills and behaviour (e.g. working memory, attention and inhibition, among others), also called executive functions (Miyake & Friedman, 2012), current knowledge about the nature of this relationship is limited. Most researchers who explore the factors conditioning successful L2 acquisition in terms of individual differences in cognitive control do not specifically address learners' pronunciation (but see Dogil & Reiterer, 2009). Moreover, research investigating L2 phonological acquisition in late learners has revealed important inter-learner variability (see also Lindemann, this volume), both in naturalistic language learning (MacKay *et al.*, 2001) and in phonetic training carried out in laboratory settings (Bradlow *et al.*, 1997; Kim & Hazan, 2010). This suggests that individual factors play an important role in the acquisition of L2 phonology, just as they play a role in other domains of L2 acquisition (Dörnyei, 2006).

Gaining a better understanding of the relationship between individual differences in cognitive control and pronunciation, especially about the role of cognitive control mechanisms and processes in the perception and production of L2 speech (e.g. phonological memory, attention, inhibition) is also of crucial importance for pronunciation assessment, both from the perspective of test takers and that of test designers and examiners. For example, learners' performance on the identification and discrimination tasks often used to assess accuracy in the perception of L2 sound contrasts is likely to be influenced by their ability to keep auditory verbal information in working memory (Cerviño-Povedano & Mora, 2011) or by their ability to focus attention on the acoustic cue that is relevant in the L2 (Safronova & Mora, 2013). Similarly, learners with stronger inhibitory control may be better able to inhibit their L1 while using the L2, which may lead to more targetlike perception and production of L2 sounds (Lev-Ari & Peperkamp, 2014). Further, perceptual judgements of nonnative speech by raters (or examiners) in terms of degree of foreign accent, comprehensibility or fluency are likely to be affected by their ability to notice, focus on and process the relevant nonnative acoustic properties in the speech signal (Isaacs & Trofimovich, 2011).

The goal of the present chapter is twofold. First, we explore the relationship between cognitive control and L2 pronunciation as a means of explaining

between-learner variation in L2 speech production accuracy. Secondly, we discuss the implications of this relationship for the assessment of L2 pronunciation. We do so in light of the findings of an empirical study investigating the role of individual differences in cognitive control in the L2 pronunciation of two groups of learners of English as a foreign language. In this study we assess the relationship between L2 learners' individual differences in cognitive control and their pronunciation accuracy, measured through acoustic analysis of L2 speech and raters' judgements of perceived comprehensibility and accentedness. Specifically, we ask whether language learners with better cognitive control are also better equipped to acquire a new phonological system as measured through pronunciation accuracy. The chapter is structured as follows. In the background section below we discuss the findings of previous research on the relationship between cognitive control and L2 phonology. We then present the empirical study and discuss the findings. Finally we present the implications of the findings with regard to the assessment of L2 pronunciation, and outline some suggestions for test design.

Background

Extensive research has uncovered some cognitive factors underlying differences in individual attainment among (mostly instructed) L2 learners. In the realm of cognitive control, working memory (Miyake & Friedman, 1998), attention (Segalowitz & Frenkiel-Fishman, 2005), and inhibitory control (Mercier *et al.*, 2014) have been associated with higher L2 proficiency or more efficient L2 processing. Regarding pronunciation abilities specifically, research also indicates that higher abilities in cognitive control relate to more accurate pronunciation and phonological processing (e.g. Darcy *et al.*, 2014, 2015; Lev-Ari & Peperkamp, 2013). However, this line of research has only evaluated isolated dimensions of phonological systems or their specific acoustic-phonetic properties (e.g. vowel categorization, voice onset time), but not global characteristics of L2 speech, such as comprehensibility or accentedness. The framework of Aptitude-Treatment Interaction (Snow, 1989) indicates that some instructional strategies (treatments) are more or less effective for particular individuals depending on their specific abilities. To our knowledge, no study has investigated the contribution of cognitive control abilities to the benefits learners receive from pronunciation instruction directly. However, indirect evidence suggests that instructional methods in which learners' attention is drawn to phonological dimensions of L2 speech enhance pronunciation improvements (e.g. Derwing *et al.*, 1998; Saito, 2011). Similarly, phonetic training studies have demonstrated improved perception and production of phonetic categories (e.g. Bradlow *et al.*, 1997). Heightened phonological awareness also relates to higher pronunciation ratings (Kennedy & Trofimovich, 2010).

These different perspectives all suggest a direct relationship between pronunciation accuracy and attention directed towards phonological dimensions of L2 speech (see also Schmidt, 1990). From there, we can infer that learners with a better ability to focus attention on speech dimensions in the input or in their own output (which roughly corresponds to the construct of noticing) and also to inhibit irrelevant information (such as interference from their L1) at the same time, might benefit more from such explicit instructional techniques (Trofimovich & Gatbonton, 2006). The present study examines how attention control, phonological short-term memory (henceforth PSTM), and inhibitory control relate to L2 learners' pronunciation accuracy. All of these cognitive abilities are implicated in the processing of L2 speech and have been operationalized through speech- or language-based measures in the present study. Attention control is the cognitive mechanism that enables individuals to flexibly and efficiently switch their focus of attention between tasks or mental sets, in this case linguistically relevant speech dimensions (see Monsell, 2003). To illustrate such a task in a learning situation, let us consider the voicing contrast in English. In obstruents (e.g. /s/ versus /z/), it is mainly cued through the duration of the preceding vowel in word-final contexts (i.e. the vowel is shorter in *place* than in *plays*) rather than the presence of voicing in the obstruent consonant (often devoiced in this context). In Spanish, by contrast, listeners attend mainly to closure voicing, so Spanish learners of English need to learn to refocus their attention onto a different dimension (duration) to properly cue the English voicing contrast in a targetlike manner.

PSTM is a short-term phonological store for verbal information which allows individuals to encode phonological units and their sequential order in the form of auditory traces that can be sustained in memory for further processing through sub-vocal rehearsal, a silent articulation mechanism (Baddeley, 2003). The use of this short-term store and the sub-vocal rehearsal mechanism are essential in learners' perception, repetition and imitation of L2 speech units and phrases.

Inhibitory control is the language control mechanism that allows bilinguals to speak one of their languages while avoiding interference from the language not in use, for example in the selection of the right word from their mental lexicon (Green, 1998). Learners with stronger inhibitory control may thus be better able to avoid interference from their L1 phonological categories and their acoustic and articulatory properties, resulting in more targetlike perception or less strongly accented L2 speech production.

The Present Study

The present study examines the relationship between three cognitive variables (attention control, PSTM and inhibitory control) and several coded

and listener-based measures of L2 pronunciation for two groups of Spanish learners of L2 English differing in their early language learning experience – monolingual Spanish speakers and bilingual Spanish-Catalan speakers. With the term 'monolingual', we refer to Spanish native speakers who grew up learning only Spanish, whereas with the term 'bilingual', we refer to native speakers of Spanish and Catalan who grew up learning both languages from an early age. Our main objective was to identify individual differences in cognitive variables relating to learners' previous bilingual experience (Bialystok, 2011) that might explain inter-learner differences in their English pronunciation skills. For attention control, we used a speeded attention task that involved switching between phonetic dimensions of auditory stimuli. We measured PSTM through a serial non-word recognition task that involved online processing of auditory verbal stimuli. Inhibitory control was measured through a lexical retrieval-induced forgetting task. Measures of L2 pronunciation accuracy included acoustic measures of an L2 vowel contrast (to determine quality and duration differences between /i:/ and /ɪ/), an accuracy measure of the production of the contrasting L2 consonants contrast /ʃ/ and /tʃ/, as well as perceptual judgements of accentedness and comprehensibility. Comprehensibility (defined as the ease/difficulty with which a listener understands accented speech) is strongly related to, yet separable from, perceived accentedness. Trofimovich and Isaacs (2012) show that, whereas accent is uniquely associated with pronunciation accuracy at the segmental, syllabic and rhythmic levels (i.e. phonology), comprehensibility is also associated with lexical richness and grammatical accuracy (see also Saito *et al.*, this volume). In the present study, a set of three sentences spoken by 51 speakers were presented to listeners. There was therefore little variation in terms of grammatical accuracy or lexical richness that could possibly affect listener ratings. Therefore, comprehensibility ratings were expected to be mainly based on pronunciation, prosody and fluency characteristics of the speech signal, which would also include voice quality and accent. The same set of sentences from all speakers was used to promote listeners' attention to segmental rather than lexical or grammatical accuracy, because opportunities for making such errors were minimized due to the constrained nature of the target materials. We included both comprehensibility and accentedness ratings as holistic measures of pronunciation in the present study under the assumption that perceived differences among L2 learners in comprehensibility and accentedness would be likely to reflect differences in the speakers' phonological processing performance (Derwing *et al.*, 2008).

We predicted that extensive long-term linguistic experience in switching between languages in a bilingual environment might provide early bilinguals with a cognitive advantage over monolinguals in cognitive control tasks measuring PSTM, attention control and inhibition. Similarly, among the early bilinguals, unbalanced bilinguals (i.e. those whose Catalan was weaker than Spanish or vice versa) were expected to outperform balanced bilinguals in

inhibitory control, as their stronger dominance in one of their two languages would require them to regularly apply a stronger level of inhibitory control to suppress the language not in use (Costa & Santesteban, 2004). We also predicted that individual differences in phonological memory, attention and inhibition, whether pertaining to individual learners' inherent capacities or driven by their bilingual experience, might provide learners with an advantage in developing more targetlike pronunciation. For example, L2 learners with larger phonological memory capacity would be better able to subvocally rehearse L2 sound sequences during speech processing, thus not only promoting the acquisition of vocabulary and grammar (e.g. French & O'Brien, 2008), but also promoting learners' capacity to notice cross-language differences between L1 and L2 sounds and between pairs of contrasting L2 sounds (e.g. MacKay *et al.*, 2001), which could enhance the formation of more accurate representations of L2 sound categories. Similarly, learners with more efficient attention control might be better able to flexibly switch their attention between L2 phonetic dimensions in contexts where they function contrastively (e.g. spectral versus duration information in vowels). And those with better inhibitory skill may be more successful at avoiding L1 phonetic interference during L2 speech production, resulting in greater segmental accuracy (Lev-Ari & Peperkamp, 2013, 2014).

Methodology

Participants

Participants included two groups of L2 learners (L1 Spanish speakers) who started learning English no sooner than schooling age (5–6) in a foreign language instructional context, with a mean age of onset (AOL) of L2 learning of 7.2 years. The monolingual group of English learners (n = 16, AOL = 7.5) grew up in a Spanish monolingual environment in Sevilla, Spain, where Spanish was the only language used at home and school and the main language they were exposed to daily through the media. Therefore, these learners' daily communication with others took place almost exclusively in Spanish. The bilingual group of English learners (n = 33, AOL = 7.0) grew up in a bilingual Spanish-Catalan environment in Barcelona, Spain. This specific language environment is characterized by a situation of intensive language contact across all kinds of communicative contexts (e.g. at home, at school, with friends, in the media), where both Spanish and Catalan are used on a daily basis by most speakers, but to varying extents depending on speakers' degree of language dominance in either Spanish or Catalan. The early bilingual participants used both Catalan and Spanish daily, but crucially differed in the amount of use of their less dominant language. Nine were relatively balanced bilinguals who

reportedly used their less dominant language (either Catalan or Spanish) 30% of the time or more on a daily basis, while 24 were unbalanced bilinguals who used their less dominant language less than 30% (13 were dominant in Spanish; 11 were dominant in Catalan). The primarily monolingual (Seville) versus bilingual (Barcelona) linguistic experience of our participant groups provides us with the ideal ground, in comparable instructional settings, for researching the effect of the interaction between cognitive skills and bilingualism on the acquisition of the pronunciation of a foreign language (English). Table 6.1 summarizes the participants' language background characteristics.

A series of Mann–Whitney *U*-Tests revealed that balanced bilinguals did not significantly differ from unbalanced bilinguals in any of the language background variables tested, except for the amount of use of their less dominant language ($U = 300$, $z = -4.37$, $p < 0.001$). Tests also showed that monolinguals differed from bilinguals only in that they were slightly older at testing ($U = 390$, $z = 2.74$, $p = 0.006$), started using the L2 at a slightly later

Table 6.1 Language background characteristics of the L2 learner groups

Measure	*Groups*					
	Monolingual (n = 16)		*Bilingual*			
			Balanced (n = 9)		*Unbalanced (n = 24)*	
	M	*SD*	*M*	*SD*	*M*	*SD*
Age at testing (years)	23.3	5.3	22.8	8.0	20.3	3.8
Age at first exposure to L2 (years)	7.5	2.1	7.5	2.6	6.8	2.6
Age at first use of L2 (years)	13.5	4.3	12.2	5.5	10.6	3.5
L2 instruction (years)	11.9	7.5	11.6	2.5	11.2	2.8
Current L2 use (0–36)[a]	17.4	5.9	15.2	5.4	16.8	5.2
Residence abroad (weeks)	5.43	10.1	2.65	5.93	6.26	21.1
Self-estimated proficiency (1–5)[b]	4.01	0.49	4.05	0.58	4.26	0.41
L2 vocabulary size (words 0–10,000)	5696	1208	6127	934	5697	982
Motivation to learn an L2 (1–9)[c]	7.6	0.8	7.7	1.2	7.8	0.6
Less dominant language (% use)	–	–	40.1	5.7	14.4	7.2

Notes: [a]This score (0–36) was obtained by adding up participants' selected level of intensity of L2 use (0 = 0%, 1 = 1–25%, 2 = 26–50%, 3 = 51–75%, 4 = 76–100%) on nine contexts of language use (e.g. with friends, at home/work, media).

[b]Participants estimated their ability to speak spontaneously, understand, read and write the L2, using five-point scales (1 = 'very poorly', 2 = 'poorly', 3 = 'passably', 4 = 'well', 5 = 'very well'). The four scores of each participant were then averaged.

[c]Average of each participant's ratings on nine motivation items presented through nine-point Likert scales (1 = 'strongly agree', 9 = 'strongly disagree'), asking participants about their motivation to learn a second language.

age ($U = 359$, $z = 2.03$, $p = 0.042$), and considered themselves slightly less motivated ($U = 163$, $z = -2.15$, $p = 0.031$). A group of native speakers of American English ($n = 10$) provided baseline data in the production task.

Procedure

The learner data reported in this chapter were collected as part of a larger project. Participants took part in a pure-tone audiometry test, an inhibitory control task (retrieval-induced forgetting), an attention control task (switching), a perception task (ABX categorization), a production task (delayed sentence repetition), and a PSTM task (serial non-word recognition). In order to control for differences in proficiency among learners, participants also took part in a L2 vocabulary size test (X-Lex/Y-Lex), because vocabulary size has been shown to be related to L2 proficiency levels (Miralpeix, 2012). Finally, they filled out a background questionnaire. Tasks were given in blocks of two or three tasks for a total testing duration of about 75–80 minutes. Task order was the same for all participants, with slight adjustments. Participants were tested individually or in groups of two in a psycholinguistics laboratory at each location. In this study we only report on the production data (perception data are reported in Darcy & Mora, 2016).

Production task

In order to elicit L2 speech productions containing the L2 contrasts of interest, we used an elicitation technique, a delayed sentence repetition task that has often been used in L2 speech production research (e.g. Trofimovich & Baker, 2006). The participants sat in a sound-isolated recording booth equipped with a microphone, headphones and a computer screen. They heard a question prompt (Voice 1), followed after 250 msec by a response (Voice 2). After a 500 msec delay, the prompt was presented again, and the participants had to repeat aloud the response heard previously. The written sentences appeared on the screen together with the first auditory presentation of the prompt/response pair, and disappeared for the second presentation of the prompt and the recording of the answer. All learners received instructions in Spanish and completed a warm-up prompt in Spanish before moving on to English. The /i:/–/ɪ/ contrast was examined together with the consonantal contrast /ʃ/–/tʃ/, because both are known to present difficulties for Spanish learners of English. The two English vowels /i:/and /ɪ/ are perceptually mapped mainly onto a single Spanish category, /i/ (Cebrian, 2006). In production, the contrast tends to be neutralized and produced as a high front vowel /i/ as well (Morrison, 2006), so that learners often implement the tense /i:/–lax /ɪ/ contrast in terms of duration (long versus short) rather than in terms of vowel quality (that is, spectral differences). The nonnative /ʃ–tʃ/ contrast has been shown to not be accurately realized in production (e.g.

Table 6.2 Sentences used to elicit the target contrasts in L2 English

Prompt		*Answer*	*Target*
/i:/–/ɪ/ contrast			
(1)	– What did he do next?	– He went by <u>ship</u> to India.	/ɪ/
(2)	– Can you see any animals in the picture?	– Only a <u>sheep</u> and a cow.	/i:/
(3)	– Why do you go jogging?	– It keeps you <u>fit</u> I think.	/ɪ/
(4)	– Shall I put the heating on?	– Yes, my <u>feet</u> are cold.	/i:/
(5)	– Do they know each other?	– Yes, they <u>sit</u> at the same table.	/ɪ/
(6)	– Is there any room left?	– Yes, there's a <u>seat</u> at the front.	/i:/
(7)	– What would you like with it?	– I'll have the <u>chips</u> please.	/ɪ/
(8)	– Did you pay a lot for it?	– No, it was a <u>cheap</u> one.	/i:/
/ʃ/–/tʃ/ contrast			
(9)	– Can you stop asking questions?	– When he <u>shows</u> us how to do it.	/ʃ/
(10)	– Did anyone get the job?	– Yes, they <u>chose</u> a brilliant person.	/tʃ/
(11)	– Did you see any animals?	– I saw a <u>sheep</u> and a horse.	/ʃ/
(12)	– Which one do you like best?	– I like the <u>cheap</u> one.	/tʃ/
(13)	– Could you buy some wine?	– All the <u>shops</u> are closed, sorry.	/ʃ/
(14)	– Are you not finishing the pork chops?	– The <u>chops</u> are too much, I'm full.	/tʃ/
(15)	– Do you get nervous at events?	– When I <u>shake</u> hands with people.	/ʃ/
(16)	– What are you looking for?	– I lost a <u>cheque</u> he gave me.	/tʃ/

Note: Target words used for accuracy measurements are underlined.

Anrrich, 2007). There were four pairs of words targeting each contrast, for a total of 16 sentences (Table 6.2). Stimuli (prompts and responses) were recorded by two female early balanced bilinguals (Mexican Spanish and American English), and were normalized for amplitude. In half the sets, one voice was used for the prompt tokens whereas the other was used for the response tokens, and the reverse was done for the remaining sets.

Two kinds of pronunciation accuracy measures were obtained from the production task: acoustic/auditory measures for the production of the target L2 English phonological contrasts (/i:/–/ɪ/ and /ʃ/–/tʃ/), and perceptual judgements of comprehensibility and accentedness provided by a panel of expert native English raters.

Acoustic/auditory measures

For the eight target words containing the /i:/ or /ɪ/ vowel, mean F1, F2 and F0 frequencies were extracted from a 15 msec window centred at the

midpoint of the steady-state portion of the second formant of the vowel. A spectral distance score, the Euclidean distance between the contrasting vowels on a Bark-normalized vowel space, was taken as a measure of accuracy in spectrally differentiating the two vowels. We predicted that the spectral distance score would be much larger for the native English control group, who would produce clearly distinct non-overlapping realizations of the /i:/ and /ɪ/ vowels, than for the L2 learners, who would produce /i:/ and /ɪ/ with varying degrees of overlap, with some learners possibly producing these two vowels identically in quality. We can infer from this measure that the smaller the spectral distance produced, the less distinctly the vowels are produced and, consequently, the harder it would be for native English listeners to perceive them as different sounds. We also computed a duration difference score in milliseconds as a measure of accuracy in temporally differentiating the two vowels. Here, a low score indicates that both vowels are of similar duration. For the /ʃ/–/tʃ/ contrast, spectrograms were visually and auditorily examined and a categorical decision was made by the researchers (and further confirmed by two naïve native listeners) about the accuracy of production. These eight target word productions, four /tʃ/-initial words and four /ʃ/-initial words (see Table 6.2), were scored as accurate if produced as palatoalveolar and exhibiting presence (for /tʃ/) or absence (for /ʃ/) of a closure. This resulted in an accuracy score out of eight.

Perceptual judgements

A panel of 20 native English listeners with a variety of L1 accents (eight American, eight British, two Irish, one Canadian, one Indian; mean age 31, range 21–48) were recruited as expert raters in order to obtain perceptual judgements of learners' pronunciation for comprehensibility (perceived difficulty in understanding) and accentedness (degree of perceived foreign accent). Raters were paid €10 for their participation and were either language teachers or language students, spoke Spanish proficiently and daily, and were also exposed to Spanish-accented English daily. They had lived in a Spanish-speaking country for an average of 31 months ($SD = 55$) and, on a series of nine-point scales, they reported being very familiar with Spanish-accented English (1 = 'not at all familiar', 9 = 'very familiar'; $M = 8.1$, $SD = 1.01$), to which they were exposed very frequently (1 = 'never', 9 = 'all the time'; $M = 7.9$, $SD = 1.02$). They also reported using Spanish frequently (1 = 'I never speak Spanish', 9 = 'I speak Spanish all the time'; $M = 6.15$, $SD = 1.92$), and to have a proficiency level in Spanish ranging from intermediate to upper-intermediate (1 = 'I can't speak Spanish', 9 = 'I speak Spanish almost like a native'; $M = 6.75$, $SD = 1.55$). In order to keep the total duration of the rating tasks no longer than 50 minutes, we selected three of the 16 English sentences, produced by each of the L2 English learners ($n = 49$). The specific

three sentences (the same for all learners) were chosen on the basis of the potential presence of a wide range of non-targetlike features for Spanish/Catalan L2 learners of English, and thus potentially also included nonnative features other than the targeted contrasts. We included the productions of six native controls (6 × 3 sentences = 18 speech samples) to motivate the use of the lower end of the accentedness (no accent) and comprehensibility (very easy to understand) scales. Raters listened to the speech samples through noise-cancelling headphones at a self-adjusted volume level in a computer-administered rating task designed using the speech analysis software, *Praat* (Boersma & Weenink, 2015). The three sentences were the responses to Prompts 5, 7 and 15 in Table 6.2. All speech samples were normalized for peak and mean amplitude. The audio files were presented in randomized order, but blocked by sentence. After a short familiarization, listeners rated them using nine-point scales, first for comprehensibility (1 = 'very easy to understand', 9 = 'very difficult to understand'), and then, in a separate task for accentedness (1 = 'no accent', 9 = 'very strong accent'). To check for rating consistency, we computed Cronbach's alpha coefficients for all speech samples ($\alpha = 0.952$ for comprehensibility and $\alpha = 0.967$ for accentedness). Given high inter-rater reliability, we computed two mean rating scores across sentences and raters for every L2 learner, one for comprehensibility and one for accentedness, and used these scores as measures of overall pronunciation accuracy.

Phonological short-term memory (PSTM) task

To measure PSTM, we used a serial non-word recognition task with Danish stimuli, so that no participant had lexical or even phonetic familiarity with the stimuli (Cerviño-Povedano & Mora, 2011). Participants were asked to identify pairs of non-word sequences of increasing length (5–6–7) as same or different. The identification accuracy of all sequence pairs (both same and different) was taken as a measure of a participant's PSTM. Table 6.3 presents two sample sequence pairs, both from illustrating different sequences.

Table 6.3 Example of two trials (length 5 and 7), both requiring the answer 'different'

Sample items	*Sequence length*						
	1	*2*	*3*	*4*	*5*	*6*	*7*
Sample 1	tys	dam	rød	mild	fup		
	tys	dam	mild	rød	fup		
Sample 2	vul	bend	sids	påk	ryd	ham	jøb
	vul	sids	bend	påk	ryd	ham	jøb

Attention control task

In this novel switching task, we used two phonological dimensions: nasality and native phonetics. All items were phonotactically legal non-words in Spanish and English. Participants had to focus their attention on either the presence of a nasal resonance in the initial segment of a non-word, or the presence/absence of an L2 accent in the pronunciation of a non-word (see Table 6.4). Participants were asked to answer one of two possible questions – 'Nasal?' versus 'Spanish?' – with respect to an auditory stimulus, by pressing two assigned buttons (yes or no) on a computer keyboard. A trial consisted of a fixation sign followed by a question displayed in the centre of the screen for 500 msec (e.g. 'Nasal?'). This question was immediately followed by an auditory stimulus (e.g. ['nofe], spoken with Spanish phonetics). Test non-words comprised 10 nasal-initial items and 10 non-nasal-initial items. Two female balanced early bilinguals (Mexican Spanish/American English) recorded both sets of stimuli, so that voice identity could not be used to determine the stimulus language.

There were two types of trials: repeat trials (R, featuring the same question as the previous trial), and switch trials (S, featuring a different question from the previous trial). Switch trials require participants to refocus their attention onto a different dimension in order to make their answer. Trials were arranged in a predictable alternating SRSR sequence (Monsell, 2003; Segalowitz & Frenkiel-Fishman, 2005). This sequence was the same for both groups. The audio tokens were randomly ordered to match the SRSR sequence; voices were randomly selected to alternate during the experiment.

Table 6.4 Non-word stimuli sample

Spanish nasal	*English nasal*	*Spanish non-nasal*	*English non-nasal*
['noma]	['noʊmə]	['piɣo]	['pʰɪgoʊ]
['nole]	['noʊleɪ]	['dofe]	['doʊfeɪ]
['niso]	['nɪsoʊ]	['saso]	['sæsoʊ]

Inhibition task

Individual differences in inhibitory control were measured as retrieval-induced inhibition (Veling & van Knippenberg, 2004), which has been shown to relate to proficient bilinguals' level of phonological interference between their languages (Lev-Ari & Peperkamp, 2013), and to intermediate learners' ability to resolve interference from their L1 during foreign-language production (Levy *et al.*, 2007). In this task, administered with E-prime in the participants' L1, participants memorized six words of three different categories (vegetables, occupations or animals) and then practised only three of the items of two categories (e.g. tomato, nurse) by typing them several times.

This increased the level of activation of the practised items and caused the inhibition of the unpractised items in these practised categories. The items from the unpractised category (i.e. animals in this case), which are not inhibited, served as control items. Participants were then tested on the recognition of all items. Those with stronger inhibition were expected to bring the unpractised items in practised categories to lower activation levels, resulting in longer retrieval reaction times (RTs) during recognition, compared to RTs for control items. An inhibitory control score was obtained by dividing the median RT of inhibited items by the median RT of control items (i.e. the higher the score above 1, the stronger the inhibition).

Vocabulary task

L2 learners' overall proficiency was measured through two receptive vocabulary size tests specially designed for L1-Spanish learners of English, X-Lex, which included English words within a 5000 frequency range (Meara & Milton, 2003), and Y-Lex, which included words within a 5000–10,000 frequency range (Meara & Miralpeix, 2007). In both tests, participants indicated by means of a mouse click whether they knew, or not, the meaning of English words appearing on the screen, including a set of control non-words. These tests provided a combined vocabulary size estimate of 0–10,000 words, which has been shown to be related to L2 proficiency levels (Miralpeix, 2012). We used this vocabulary-size measure to control for between-learner differences in L2 proficiency.

Results

Production accuracy: L2 learners versus English controls

Data from four learners obtaining scores above 2.5 standard deviations from the mean suggesting unusual difficulty or atypical performance on the task (one in the inhibition task, and three in the attention control task) were excluded from the analyses in order to achieve normality in the distribution of the scores, leaving a total of 15 monolingual and 30 bilingual (nine balanced, 21 unbalanced) learners for analysis. However, for the comparisons between native English controls and L2 learners we used non-parametric Mann–Whitney *U*-Tests, as there was a large difference in sample size between the groups and the distribution of the native English control group scores was skewed. As expected, native English controls produced the /i:/–/ɪ/ contrast with significantly larger spectral distances ($M = 3.87$) than L2 learners did ($M = 0.59$) $U = 0.000$, $z = -4.91$, $p < 0.001$ (see Table 6.5a). L2 learners' duration difference score ($M = 1.67$) was also significantly smaller than that of native controls ($M = 3.51$) $U = 129$, $z = -2.09$, $p = 0.036$, suggesting that unlike native controls they could not distinguish /i:/ from /ɪ/ in either quality or duration.

Table 6.5a Pronunciation accuracy and vocabulary size for the learner and control groups

Measure	*Groups*					
	Monolingual (n = 15)		*Bilingual (n = 30)*		*Native controls (n = 10)*	
	M	*SD*	*M*	*SD*	*M*	*SD*
Vowel spectral distance (Bark)	0.71	0.37	0.53	0.39	3.87	0.86
Vowel duration difference (msec)	0.43	0.47	2.30	15.4	3.51	0.78
Consonant accuracy (0–8)	6.97	1.29	7.60	1.07	8.00	0.00
Comprehensibility (1–9)	3.97	0.88	3.60	0.56	1.34	0.13
Accentedness (1–9)	5.45	1.15	5.20	0.74	1.23	0.10
Vocabulary size (0–10,000)	5753	1228	5796	1015	–	–

The /ʃ/–/tʃ/ contrast was produced at very high accuracy rate by L2 learners ($M = 7.4$ out of a maximum score of 8.0), which did not differ significantly from that of native controls ($M = 8.0$), suggesting that this consonant contrast did not present difficulties for the L2 learners (see Table 6.5).

As regards native English raters' assessment of overall pronunciation accuracy, the ratings of the speech samples by the six native English controls yielded scores at the lowest end of the scales for both comprehensibility ($M = 1.34$, $SD = 0.13$) and accentedness ($M = 1.22$, $SD = 0.09$), suggesting that these productions were perceived to be at the high end of ability (i.e. very easy to understand and free of a foreign accent, respectively). In

Table 6.5b Pronunciation accuracy and vocabulary size for the bilinguals (balanced vs. unbalanced)

Measure	*Groups*			
	Balanced bilinguals (n = 9)		*Unbalanced bilinguals (n = 21)*	
	M	*SD*	*M*	*SD*
Vowel spectral distance (Bark)	0.57	0.34	0.51	0.42
Vowel duration difference (msec)	4.46	15.9	1.37	15.5
Consonant accuracy (0–8)	7.89	0.33	7.48	1.25
Comprehensibility (1–9)	3.57	0.46	3.61	0.61
Accentedness (1–9)	5.22	0.52	5.19	0.83
Vocabulary size (0–10,000)	6127	934	5654	1037

Notes: Raters provided perceptual assessments of overall pronunciation accuracy on nine-point scales for comprehensibility (1 = very easy to understand, 9 = very difficult to understand) and accentedness (1 = no accent, 9 = very strong accent).

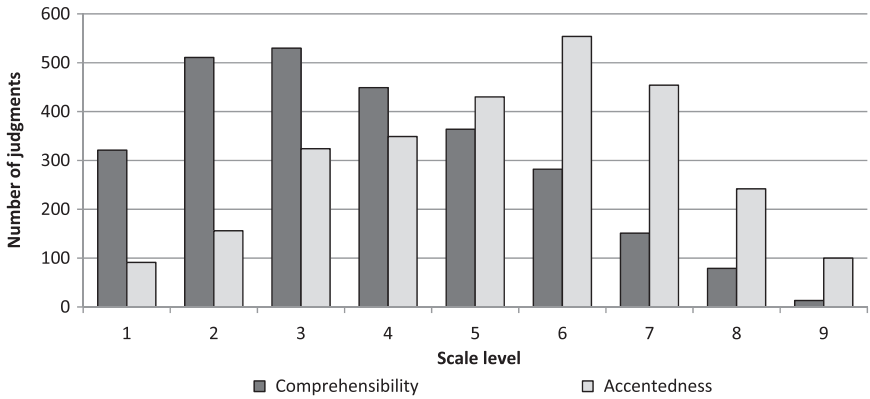


Figure 6.1 Histogram illustrating the total number of judgements at each scale level for L2 learners' speech samples by dimension (3 sentences × 45 learners × 20 raters = 2700)

Notes: 1 = 'very easy to understand/no accent'; 9 = 'very difficult to understand/very strong accent'.

contrast to the ratings for this cohort, which concentrated at the positive ends of the scales, the ratings for L2 learners' speech samples spanned the entire scale. They were lower for comprehensibility ($M = 3.72$, $SD = 0.68$) than for accentedness ($M = 5.28$, $SD = 0.89$), suggesting that despite being relatively strongly accented, the samples were overall quite easy to understand (see Figure 6.1).

Comprehensibility and accentedness ratings were strongly related to one another ($r = 0.899$, $p < 0.001$), indicating that, given the short length of the speech materials the judges were asked to rate, difficulty in understanding was associated mainly to nonnative segmental features, so that L2 learners with heavier accents were also perceived to be harder to understand. However, acoustic measures of pronunciation accuracy for the vowels were unrelated to either ratings (all rs < -0.1, $p > 0.05$), but the consonant production accuracy scores were ($r = -0.575$, $p < 0.001$ and $r = -0.509$, $p < 0.001$, respectively), suggesting that the raters might have been more sensitive to the inaccurate production of the /ʃ/–/tʃ/ contrast (substituting /ʃ/ for /s/ or /tʃ/) than to the lack of a spectral distinction between /i:/ and /ɪ/.

Finally, as previous research has shown that vocabulary size is related to accuracy in L2 vowel perception (Bundgaard-Nielsen *et al.*, 2011), we explored the relationship between L2 learners' vowel accuracy in production and their vocabulary size. L2 learners had on average a medium vocabulary size corresponding to an intermediate level of English proficiency ($M = 5782$, $SD = 1077$), but there was considerable between-learner variation in the scores, which ranged from 3200 words for the lowest score to 8450 for the

highest (range: 5250). Vocabulary size was related to higher vowel production accuracy (spectral distance score: $r = 0.356$, $p = 0.017$; duration difference score: $r = 0.291$, $p = 0.052$), to better comprehensibility ($r = -0.272$, $p = 0.071$) and to lower accentedness ($r = -0.320$, $p = 0.032$), indicating that more proficient learners (those with larger vocabularies) are also those with more accurate L2 pronunciation.

Production accuracy: L2 learner groups

Monolingual L2 learners realized the /i:/–/ɪ/ contrast with a slightly larger spectral distance and a much shorter duration difference than bilingual L2 learners did, and also produced the /ʃ/–/tʃ/ less accurately than bilinguals. They were also perceived to be slightly more comprehensible and to have slightly milder accents than bilinguals, but appeared to have very similar vocabulary sizes (see Table 6.5a). However, none of these differences, summarized in Table 6.6, reached significance. A similar result was obtained for the mean pronunciation accuracy of balanced and unbalanced bilinguals, except for the duration difference score, which was much higher in balanced than in unbalanced bilinguals. Balanced bilinguals also appeared to have a slightly larger vocabulary size than monolinguals. None of these differences reached significance either. Thus, differences between the monolingual and bilingual L2 learner groups, and between the balanced and unbalanced bilingual groups, were small and the groups were comparable in L2 pronunciation accuracy.

Table 6.6 Matched-pairs *t*-tests comparing L2 learner groups

Measure	*Monolingual vs. bilingual*		*Balanced vs. unbalanced*	
	t(43)	*p*	*t(28)*	*p*
Spectral distance	–1.52	0.14	0.42	0.68
Duration difference	0.66	0.51	0.50	0.62
Consonant accuracy	1.75	0.09	1.40	0.17
Comprehensibility	–1.73	0.09	–0.19	0.85
Accentedness	–0.88	0.39	0.09	0.93
Vocabulary size	0.13	0.90	0.18	0.25

Cognitive tasks

L2 learners' scores on PSTM, attention control and inhibition were unrelated to one another for both monolingual and bilingual groups (all $r < 0.25$ and $p > 0.15$), confirming that they were measuring three different constructs. Differences between monolingual and bilingual L2 learners were found on attention and inhibitory control. Against our predictions, the

Table 6.7 Descriptive statistics for the cognitive tasks for the L2 learner groups

	Monolingual		*Bilingual*		*Balanced bilinguals*		*Unbalanced bilinguals*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
PSTM (% correct)	64.68	14.41	63.68	11.99	62.11	8.53	64.35	13.33
Attention control shift cost (msec)	48.84	59.98	59.87	54.31	59.79	49.24	59.91	57.50
Inhibitory control (ratio)	1.05	0.29	1.11	0.28	0.97	0.23	1.18	0.29

monolingual group outperformed the bilingual group on attention control; that is, the cost of shifting between phonetic dimensions was smaller for the monolingual L2 learners than it was for the bilingual group. However, large variability associated with this measure (see means and standard deviations in Table 6.7) renders the group differences negligible and underlines considerable individual differences in inhibitory control in both groups. The bilingual group outperformed the monolingual group on inhibitory control, obtaining a higher mean score on this measure. For PSTM, all groups performed similarly but, again, there was considerable variation within groups.

A series of independent-samples *t*-tests confirmed that the monolingual group did not significantly differ from bilinguals in attention control ($t(43) = 0.621$, $p = 0.538$), inhibitory control ($t(43) = 0.663$, $p = 0.511$) or PSTM ($t(43) = -0.245$, $p = 0.807$). Within the bilingual group, unbalanced bilinguals outperformed balanced bilinguals on inhibitory control to a marginally significant level ($t(28) = -1.896$, $p = 0.068$), but both bilingual groups performed similarly in attention control ($t(28) = -0.006$, $p = 0.996$) and PSTM ($t(28) = -0.462$, $p = 0.648$).

Relationship between cognitive skills and pronunciation accuracy

In order to assess the extent to which between-learner variation in pronunciation accuracy can be attributed to individual differences in cognitive ability, we explored the relationship between pronunciation accuracy and cognitive ability scores through a regression analysis. Because vocabulary size was found to be significantly associated with L2 learners' production accuracy, we used a hierarchical regression controlling for vocabulary size as a means of isolating the contribution of individual differences in PSTM, attention and inhibition to explain between-learner variation in pronunciation accuracy. These analyses were carried out separately for the monolingual ($n = 15$) and the bilingual ($n = 30$) groups.

The hierarchical regression analysis (see Appendix to this chapter) using vocabulary size scores to partial out inter-learner differences in proficiency

showed that attention and PSTM contributed significantly to pronunciation accuracy scores only for monolingual L2 learners, accounting for 38.9% ($p = 0.016$) and 25% ($p = 0.016$) of the variance, respectively, in the duration difference scores. This suggests that monolingual learners with better phonological memory capacity and stronger attention control produced a larger, more targetlike duration difference between /i:/ and /ɪ/. For this group, attention was also found to significantly account for 44.1% ($p = 0.010$) of the variance in the spectral distance scores. However, the direction of this relationship was contrary to what we had predicted; that is, monolingual L2 learners who found it harder to switch between dimensions in the attention switching task (i.e. those with higher attention shift costs) produced the /i:/–/ɪ/ contrast more accurately, with larger quality difference (see below for discussion). Cognitive variables were not found to contribute significantly to pronunciation accuracy scores in the early bilingual L2 learner group after controlling for differences due to vocabulary size.

Discussion and Conclusion

This chapter has explored the relationship between cognitive control and L2 pronunciation accuracy. We set out with the prediction that between-learner variation in pronunciation accuracy, measured both instrumentally through acoustic analysis and holistically through comprehensibility and accentedness ratings, could partly be accounted for by individual differences in cognitive control. However, the pattern of results emerging from the present study appears to be rather complex, as the relationship between individual differences in the three cognitive skills examined (PSTM, attention and inhibition) and pronunciation accuracy is not observable generally across all pronunciation measures and for all L2 learner groups and, when present, it is not always in the expected direction.

For example, it was only for monolingual learners that attention control was found to be strongly related to vowel production accuracy, but the relationship was such that learners who were less efficient at switching between dimensions in the attention task (those with larger switching costs) were those who were more accurate at producing the /i:/–/ɪ/ contrast in the delayed picture naming task. This apparently puzzling relationship between less efficient attention control and higher pronunciation accuracy can be explained by the non-unitary nature of attention as a phenomenon (Cohen *et al.*, 2004; Tomlin & Villa, 1994), which may involve various cognitive processes and mechanisms (e.g. selection, switching and inhibition, among others), and by what a switching task and a delayed sentence repetition task require in terms of the recruitment of learners' attentional resources during task performance. Whereas in the attention switching task learners have to flexibly and rapidly reallocate their attention on alternating dimensions

(Monsell, 2003), in the delayed sentence repetition task learners had to focus their attention on the target sentence so that it could be accurately repeated after an intervening sentence. Previous research has shown that different cognitive mechanisms underlie selective attention and attention switching (Fan *et al.*, 2005). Thus, learners with stronger 'selective' attention may have been better able to focus on the L2 target sentences (or the target words in those sentences) for delayed repetition, while at the same time performing less efficiently in an auditory attention task that required 'switching' rapidly between speech dimensions (see also Koch *et al.*, 2011). The fact that PSTM was significantly associated with larger /i:/–/ɪ/ duration differences is also consistent with the nature of the delayed sentence repetition task: learners with larger PSTM capacity would be able to more efficiently process and sub-vocally rehearse the target sentences before repetition, which would result in more accurate pronunciation.

However, none of the relationships found between cognitive control (attention and PSTM) and pronunciation accuracy for the monolingual learner group was observed for the bilingual group. Why did individual differences in cognitive control, also present in the bilingual learner group, fail to show any systematic relationship with pronunciation accuracy measures? What makes an extensive bilingual linguistic experience 'different'? We propose that bilinguals' extensive practice in switching between languages and in the daily use of two language systems might make attention and PSTM less crucial to the development of L2 pronunciation accuracy than in the case of a monolingual L2 learner, for whom individual differences in cognitive control might play a more fundamental role in L2 phonological development. That is, L2 learners' previous linguistic experience as either 'monolingual' or 'bilingual' might mediate the role of cognitive skills in shaping learners' L2 phonological development. Although bilingualism may provide individuals with a general advantage in cognitive control (Bialystok, 2011), the bilingual experience in an environment like Barcelona where the two languages are constantly used (as opposed to monolingualism) also provides individuals with a rich extensive language contact experience across a whole range of language use patterns. Thus, for example, balanced bilinguals with poor attention control may have performed well in our attention control task due to their highly trained expertise at switching between languages, compensating for their perhaps initially limited attention control ability. Similarly, unbalanced bilinguals with initially poor inhibitory control may have performed well in our inhibitory control task because the unbalanced pattern of language use in their daily lives requires them to apply stronger inhibition when using the language they speak less frequently, thus training their inhibitory control in general. Consequently, the bilingual group displays individual variation due to their bilingual experience which might affect their performance on our cognitive tasks, and add to their individual differences in cognitive control. However, for monolinguals, individual differences in cognitive control are not

confounded with experiential factors and have a better chance of being more directly related to individual variation in L2 pronunciation performance.

Indirect evidence indicates that instructional methods in which learners' attention is drawn to phonological dimensions, such as explicit pronunciation instruction or corrective feedback, enhance pronunciation improvement. This suggests a close relationship between pronunciation accuracy and directing attention towards phonological dimensions, at least for those learners without prior L2 learning experience. Learners with a better ability to focus attention on speech dimensions in the input or in their own productions might benefit more from explicit pronunciation instruction techniques than learners with poorer attention control (Saito, 2013; Saito & Lyster, 2012). Clearly, in order to gain a better understanding of the role of cognitive control on pronunciation development, much empirical research is needed in this area. A further complication is the variety of tasks employed to measure PSTM, attention and inhibition. Future research should establish the basis for using speech-based or domain-general (i.e. non-linguistic) cognitive tasks capable of capturing more precisely those individual differences in the recruitment of cognitive resources during L2 phonological processing that would have a more direct impact on L2 pronunciation performance.

Implications

A crucial question to address is how our findings relate to assessing L2 pronunciation. We see potential consequences of our findings for pronunciation assessment in three main areas. First, our findings indicate that the specific type of task and its cognitive demands might advantage certain groups of learners. For example, the task may lead certain learners (with higher PSTM, attention, etc.) to obtain better performance for reasons that are unrelated to their phonological skill/pronunciation ability. An ideal solution to this problem could be to first test each learner on a PSTM task in order to assign this learner to a task that has been adjusted for various ranges of PSTM ability. Short of having such tasks and time at their disposal, a more equitable and feasible approach would be likely to entail using various kinds of tasks that would help compensate for individual differences in cognitive control. To illustrate this, using tasks such as sentence repetition (with higher PSTM demand) along with picture naming (with lower PSTM demand) may be a valid option so as not to disadvantage learners with smaller working memory capacity. Similarly, using vocabulary that is already acquired and familiar to students to assess pronunciation could be combined with short non-word repetition of up to four syllables. This approach would equally suit learners with large or small vocabularies and with large or small PSTM, and possibly reduce performance discrepancies that are due to cognitive control differences. Such tasks (sentence repetition, picture naming,

non-word repetition, etc.) are well suited for assessment types or diagnostic purposes that require controlled speech materials, in order to examine whether learners have successfully acquired word-initial onset clusters or specific consonants, for example. Their disadvantage is that they do not clearly relate to what learners need to do in real-world settings. For assessing pronunciation on larger speech chunks or on more authentic task types (such as a short oral summary of an event, or an oral presentation), a common assessment tool is an evaluation by various raters of a sample of extemporaneous speech. Here also it is important to consider the possible consequences of task demand and individual differences in cognitive control. Perhaps allowing learners to rehearse the content of an answer or to practise a lecture extract, for example, may limit the incidence of attention control or PSTM constraints on production, without unduly modifying individual speech patterns. A related concern is the item and/or sentence complexity used in a task – especially if it contains unfamiliar words or structures – which might disproportionately affect certain learners with less efficient cognitive control and negatively affect their pronunciation performance.

The second area for which our findings are important is the kind of switching required by the task or the task sequence. For example, some learners may perform less well on language switching tasks such as oral translation tasks for reasons more related to attention control or inhibition than their pronunciation. This will also have an impact on different learning contexts in different ways, for example, as it is contingent on the language of instruction. In immersion contexts such as intensive English programmes in the United States, the language of instruction is in the vast majority of cases the same as the one of assessment, which reduces the likelihood of switching. The situation may be different in other contexts, such as a bilingual context or a foreign-language classroom, where it is possible that the language spoken in the classroom is not always the language being learned. For instance, the instruction for an oral task in the L2 may be given in the L1 of the learners. While it remains unclear how much the interaction between task and context exactly impacts pronunciation scores, it is possible that a lesser amount of switching is more favourable to certain learners over others and, as a result, makes it easier for some to inhibit their L1 during an L2 task. While this may not have a strong effect on scores within a single classroom, it may be of importance for test makers who are designing pronunciation assessment tools as a part of standardized tests. This is particularly the case when production accuracy is an assessed component, including using automated oral assessments that rely on measures such as pause length and placement, words stress and segmental accuracy to derive machine-generated scores.

Thirdly, while not directly relevant for the kind of assessment such as a final oral examination, but very important in the classroom for ongoing assessment, our findings also suggest that noticing abilities may very well vary among learners as a function of cognitive control, which in turn may

have an impact on pronunciation progress. Consequently, we surmise that it might be important to use various types of explicit feedback strategies to compensate for potential differences here as well.

We have outlined some suggestions for test design, but it is at present unclear to what extent these aspects of task complexity or task design will measurably affect pronunciation assessment outcomes. There is clearly a need for targeted research in this area.

Finally, one possible long-term application of this research is to contribute to creating usable professional knowledge for teachers in adult L2 pronunciation instruction, and to develop more efficient pronunciation instruction methods rooted in psycholinguistic research on the role of cognitive control in L2 pronunciation. In order to reach this goal, it is necessary first to understand the extent to which cognitive control impacts on L2 learners' pronunciation improvements in addition to elucidating the cognitive abilities that most contribute to successful pronunciation learning. Future research should also empirically evaluate the benefits of developing new cognitive training methods and technologies to enhance pronunciation instruction (Jaeggi *et al.*, 2011) – for example, through training selective attention (Tang & Posner, 2009). Training cognitive skills underlying phonological processing will have important implications not only for learners' development of L2 pronunciation, but also for the testers' assessment of nonnative speech. Our own contribution has been exploratory, laboratory-based and mostly correlational in nature. At present, an *explanatory* account of the role of individual differences in cognitive control in L2 pronunciation learning has yet to be conducted, and remains a fascinating avenue for future research.

Acknowledgements

We would like to thank Paola Rodrigues, Tanya Flores, Maggie Peters and Fiona Pannett for recording the stimuli, and Elena Safronova, Eva Cerviño-Povedano, Marina Barrio Parra and M. Heliodora Cuenca Villarín for help with testing and data processing in Seville and Barcelona. We are indebted to Carmen Muñoz and Kathleen Bardovi-Harlig for institutional and financial help. We further acknowledge the following grant support: Grant-in-Aid, Indiana University Bloomington; Grants FFI2013-47616-P (Ministerio de Economía y Competitividad), 2014SGR1089 (Generalitat de Catalunya), and *Language Learning's Small Grants Research Program 2013*.

References

Abrahamsson, N. and Hyltenstam, K. (2009) Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning* 59 (2), 249–306.

Anrrich, G.M. (2007) Substitutions for English consonants by adult speakers of Cuban Spanish. Unpublished doctoral dissertation, Georgetown University, Washington, DC. Available from ProQuest Dissertations and Theses database (UMI No. 3302093).

Baddeley, A.D. (2003) Working memory and language: An overview. *Journal of Communication Disorders* 36, 189–208.

Bialystok, E. (2011) Reshaping the mind: The benefits of bilingualism. *Canadian Journal of Experimental Psychology* 65, 229–235.

Boersma, P. and Weenink, D. (2015) *Praat: doing phonetics by computer* [Computer program]. Version 5.4.08, retrieved 25 March 2015 from http://www.praat.org/

Bradlow, A.R., Pisoni, D.B., Akahane-Yamada, R. and Tohkura, Y.I. (1997) Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America* 101, 2299–2310.

Bundgaard-Nielsen, R.L., Best, C.T. and Tyler, M.D. (2011) Vocabulary size matters: The assimilation of second-language Australian English vowels to first-language Japanese vowel categories. *Applied Psycholinguistics* 32, 51–67.

Cebrian, J. (2006) Experience and the use of duration in the categorization of L2 vowels. *Journal of Phonetics* 34, 372–387.

Cerviño-Povedano, E. and Mora, J.C. (2011) Investigating Catalan learners of English over-reliance on duration: Vowel cue weighting and phonological short-term memory. In K. Dziubalska-Kołaczyk, M. Wrembel and M. Kul (eds) *Achievements and Perspectives in The Acquisition of Second Language Speech: New Sounds 2010* (pp. 53–64). Frankfurt am Main: Peter Lang.

Cohen, J.D., Aston-Jones, G. and Gilzenrat, M.S. (2004) A systems-level perspective on attention and cognitive control: Guided activation, adaptive gating, conflict monitoring, and exploitation versus exploration. In M.I. Posner (ed.) *Cognitive Neuroscience of Attention* (pp. 71–90). New York: Guilford Press.

Costa, A. and Santesteban, M. (2004) Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of Memory and Language* 50, 491–511.

Darcy, I. and Mora, J.C. (2016) Executive control and phonological processing in Spanish learners of English: The effect of learning in a monolingual or a bilingual context. In G. Granena, D.O. Jackson and Y. Yilmaz (eds) *Cognitive Individual Differences in L2 Processing and Acquisition* (pp. 247–275). Bilingual Processing and Acquisition series. Amsterdam: John Benjamins.

Darcy, I., Ewert, D. and Lidster, R. (2012) Bringing pronunciation instruction back into the classroom: An ESL teachers' pronunciation 'toolbox'. In J. Levis and K. LeVelle (eds) *Proceedings of the 3rd Pronunciation in Second Language Learning and Teaching Conference* (pp. 93–108). Ames, IA: Iowa State University.

Darcy, I., Mora, J.C. and Daidone, D. (2014) Attention control and inhibition influence phonological development in a second language. *Concordia Working Papers in Applied Linguistics* 5, 115–129.

Darcy, I., Park, H. and Yang, C.L. (2015) Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learning and Individual Differences* 40, 63–72.

Derwing, T.M., Munro, M.J. and Wiebe, G. (1998) Evidence in favor of a broad framework for pronunciation instruction. *Language Learning* 48, 393–410.

Derwing, T.M., Munro, M.J. and Thomson, R.I. (2008) A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics* 29, 359–380.

Dogil, G. and Reiterer, S.M. (eds) (2009) *Language Talent and Brain Activity*. Berlin: Mouton de Gruyter.

Dörnyei, Z. (2006) *The Psychology of the Language Learner: Individual Differences in Second Language Acquisition*. New York: Routledge.

Fan, J., McCandliss, B.D., Fossella, J., Flombaum, J.I. and Posner, M.I. (2005) The activation of attentional networks. *NeuroImage* 26 (2), 471–479.

French, L.M. and O'Brien, I. (2008) Phonological memory and children's second language grammar learning. *Applied Psycholinguistics* 29, 463–487.

Gluszek, A. and Dovidio, J.F. (2010) The way they speak: A social psychological perspective on the stigma of non-native accents in communication. *Personality and Social Psychology Review* 14, 214–237.

Golestani, N. and Zatorre, R.J. (2009) Individual differences in the acquisition of second language phonology. *Brain and Language* 109 (2), 55–67.

Green, D.W. (1998) Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition* 1, 67–81.

Hu, X., Ackermann, H., Martin, J.A., Erb, M., Winkler, S. and Reiterer, S.M. (2013) Language aptitude for pronunciation in advanced second language (L2) learners: Behavioural predictors and neural substrates. *Brain and Language* 127 (3), 366–376.

Isaacs, T. and Trofimovich, P. (2011) Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics* 32 (1), 113–140.

Jaeggi, S.M., Buschkuehl, M., Jonides, J. and Shah, P. (2011) Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences* 108, 10081–10086.

Kennedy, S. and Trofimovich, P. (2010) Language awareness and second language pronunciation: A classroom study. *Language Awareness* 19, 171–185.

Kim, Y.H. and Hazan, V. (2010) Individual variability in the perceptual learning of L2 speech sounds and its cognitive correlates. In K. Dziubalska-Kołaczyk, M. Wrembel and M. Kul (eds) *New Sounds 2010: Proceedings of the 6th International Symposium on the Acquisition of Second Language Speech* (pp. 251–256). Frankfurt am Main: Peter Lang.

Koch, I., Lawo, V., Fels, J. and Vorländer, M. (2011) Switching in the cocktail party: Exploring intentional control of auditory selective attention. *Journal of Experimental Psychology: Human Perception and Performance* 37, 1140–1147.

Lev-Ari, S. and Peperkamp, S. (2013) Low inhibitory skill leads to non-native perception and production in bilinguals' native language. *Journal of Phonetics* 41, 320–331.

Lev-Ari, S. and Peperkamp, S. (2014) The influence of inhibitory skill on phonological representations in production and perception. *Journal of Phonetics* 47, 36–46.

Levy, B.J., McVeigh, N.D., Marful, A. and Anderson, M.C. (2007) Inhibiting your native language: The role of retrieval-induced forgetting during second-language acquisition. *Psychological Science* 18, 29–34.

MacKay, I.R.A., Meador, D. and Flege, J.E. (2001) The identification of English consonants by native speakers of Italian. *Phonetica* 58, 103–125.

Meara, P. and Milton, J. (2003) X_Lex, *The Swansea Levels Test.* Newbury: Express Publishing.

Meara, P.M. and Miralpeix, I. (2007) *D_Tools* (Version 2.0; _lognostics: Tools for vocabulary researchers: lognostics) [Computer software]. Swansea: University of Wales. See http://www.lognostics.co.uk/tools/index.htm (accessed 30 March 2012).

Mercier, J., Pivneva, I. and Titone, D. (2014) Individual differences in inhibitory control relate to bilingual spoken word processing. *Bilingualism: Language and Cognition* 17 (1), 89–117.

Miralpeix, I. (2012) X_Lex and Y_Lex: A validation study. *22nd VARG Conference*. Newtown: Vocabulary Acquisition Research Group.

Miyake, A. and Friedman, N.P. (1998) Individual differences in second language proficiency: Working memory as language aptitude. In A.F. Healy and L.E. Bourne (eds) *Foreign Language Learning: Psycholinguistic Studies on Training and Retention* (pp. 339–364). Mahwah, NJ: Lawrence Erlbaum.

Miyake, A. and Friedman, N.P. (2012) The nature and organization of individual differences in executive functions four general conclusions. *Current Directions in Psychological Science* 21 (1), 8–14.

Monsell, S. (2003) Task switching. *Trends in Cognitive Sciences* 7, 134–140.
Morrison, G.S. (2006) L1 & L2 production and perception of English and Spanish vowels: A statistical modelling approach. Unpublished doctoral dissertation, University of Alberta.
Moyer, A. (1999) Ultimate attainment in L2 phonology. *Studies in Second Language Acquisition* 21 (1), 81–108.
Safronova, E. and Mora, J.C. (2013) Attention control in L2 phonological acquisition. In A. Llanes Baró, L. Astrid Ciro, L. Gallego Balsà and R.M. Mateus Serra (eds) *Applied Linguistics in the Age of Globalization* (pp. 384–390). Lleida: Edicions de la Universitat de Lleida.
Saito, K. (2011) Examining the role of explicit phonetic instruction in native-like and comprehensible pronunciation development: An instructed SLA approach to L2 phonology. *Language Awareness* 20, 45–59.
Saito, K. (2013) Re-examining effects of form-focused instruction on L2 pronunciation development: The role of explicit phonetic information. *Studies in Second Language Acquisition* 35, 1–29.
Saito, K. and Lyster, R. (2012) Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /r/ by Japanese learners of English. *Language Learning* 62, 595–633.
Schmidt, R.W. (1990) The role of consciousness in second language learning. *Applied Linguistics* 11, 129–158.
Segalowitz, N. and Frenkiel-Fishman, S. (2005) Attention control and ability level in a complex cognitive skill: Attention shifting and second-language proficiency. *Memory and Cognition* 33, 644–653.
Snow, R. (1989) Aptitude-treatment interaction as a framework for research on individual differences in learning. In P. Ackerman, R.J. Sternberg and R. Glaser (eds) *Learning and Individual Differences* (pp. 13–59). New York: W.H. Freeman.
Tang, Y.Y. and Posner, M.I. (2009) Attention training and attention state training. *Trends in Cognitive Sciences* 13, 222–227.
Thomson, R.I. and Derwing, T.M. (2015) The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics* 36 (3), 326–344.
Tomlin, R.S. and Villa, V. (1994) Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition* 16, 183–203.
Trofimovich, P. and Baker, W. (2006) Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition* 28 (1), 1–30.
Trofimovich, P. and Gatbonton, E. (2006) Repetition and focus on form in processing L2 Spanish words: Implications for pronunciation instruction. *Modern Language Journal* 90, 519–535.
Trofimovich, P. and Isaacs, T. (2012) Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition* 15, 905–916.
Veling, H. and van Knippenberg, A. (2004) Remembering can cause inhibition: Retrieval-induced inhibition as cue independent process. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30, 315–318.

Appendix: Results of a Hierarchical Multiple Regression Analysis Using Attention and PSTM as Predictors of Pronunciation Accuracy Scores

Group	*Criterion variable*	*Predictor*	R^2	ΔR^2	*B*	*95% CI for B*	*t*	*p*
Bilingual	Spectral distance	Vocabulary size	0.330	0.330	0.574	[0.000,0.000]	3.713	0.001
		Attention	0.343	0.013	0.113	[−0.002,0.003]	0.721	0.477
		PSTM	0.391	0.048	0.232	[−0.003,0.019]	1.434	0.164
	Duration difference	Vocabulary size	0.143	0.143	0.378	[0.000,0.011]	2.163	0.039
		Attention	0.151	0.007	0.087	[−0.079,0.128]	0.487	0.630
		PSTM	0.236	0.085	−0.308	[−0.877,0.082]	−1.705	0.100
	Comprehensibility	Vocabulary size	0.082	0.082	−0.286	[0.000,0.000]	−1.582	0.125
		Attention	0.199	0.117	0.342	[0.000,0.007]	1.985	0.057
		PSTM	0.284	0.085	−0.308	[−0.031,0.019]	−1.758	0.090
	Accentedness	Vocabulary size	0.082	0.082	−0.286	[0.000,0.000]	−1.579	0.126
		Attention	0.114	0.032	0.178	[−0.003,0.008]	0.984	0.334
		PSTM	0.170	0.057	−0.251	[−0.040,0.008]	−1.331	0.195
Monolingual	Spectral distance	Vocabulary size	0.000	0.000	0.005	[0.000,0.000]	0.019	0.985
		Attention	0.441	0.441	0.667	[0.001,0.009]	3.077	0.010
		PSTM	0.511	0.070	0.264	[0.004,0.029]	1.254	0.236
	Duration difference	Vocabulary size	0.010	0.010	0.098	[0.000,0.000]	0.356	0.728
		Attention	0.399	0.389	0.627	[0.001,0.009]	2.787	0.016
		PSTM	0.649	0.250	0.500	[0.004,0.0029]	2.802	0.017
	Comprehensibility	Vocabulary size	0.073	0.073	−0.270	[−0.001,0.000]	−1.092	0.298
		Attention	0.117	0.044	0.211	[−0.006,0.012]	0.824	0.428
		PSTM	0.231	0.114	−0.337	[−0.056,0.015]	−1.275	0.228
	Accentedness	Vocabulary size	0.133	0.133	−0.364	[−0.001,0.000]	−1.410	0.182
		Attention	0.189	0.057	0.239	[−0.006,0.016]	0.917	0.377
		PSTM	0.319	0.129	−0.360	[−0.072,0.015]	−1.446	0.176

7 Students' Attitudes Towards English Teachers' Accents: The Interplay of Accent Familiarity, Comprehensibility, Intelligibility, Perceived Native Speaker Status, and Acceptability as a Teacher

Laura Ballard and Paula Winke

Introduction

This chapter sets out to discuss nonnative-English speaking students' perceptions of their nonnative-English speaking teachers and whether those perceptions are based on the teachers' accents. In particular, the study focuses on the relationship between a teacher's accentedness, a student's understanding of the teacher's accented speech, and the student's subsequent perception of the teacher's acceptability as an English language teacher. Unlike some of the other chapters in this book, we do not directly examine students' acquisition of pronunciation or accents; rather, we delve into students' underlying perceptions based on accent. Understanding this area of pronunciation and accents in language learning programmes is important because underneath any desire to learn or acquire a specific type of pronunciation is a certain motivation to do so. And those desires and motivations may be tied to the students' perceptions of the world. By looking within this specific context, have a lens into the language students' accent worldview.

Background

In general, if you ask any English as a second language learner what level of a *foreign* accent they wish to have when they speak English, they will say none. Their ultimate goal most likely will be 'nativelike, accent-free pronunciation' (Crowther *et al.*, 2015: 81; see also Scales *et al.*, 2006). The learners may want a nativelike accent because it is considered prestigious (see Davies, this volume; Tokumoto & Shibata, 2011) or necessary to avoid social and workplace discrimination (Lippi-Green, 2012). Applied linguists, however, will most likely argue that what is important is not a nativelike accent, but rather being easily understood (see Derwing & Munro, 2009; Harding, this volume; Levis, 2005).

Indeed, language learners' desires to sound nativelike is, perhaps, closely related to their wanting to be easily understood or, as Gluszek and Dovidio (2010) reported, related to wanting to be included, accepted or perceived as a member of the target social group. Learners themselves may not be able to articulate the differences between not having an accent and being easily understood. Scales *et al.* (2006: 735) found that a majority of the language learners in their study wanted to sound like a native speaker, but few could actually identify whether someone was a native or nonnative speaker. Some learners stated that they wanted a native accent for future employability, but the majority expressed that 'the native accent was the obvious choice to strive for' without being able to give concrete reasons why. Learners may view a nativelike accent as the means to acquire their real goals of acceptance and equality, but it is not certain that a nativelike accent is indeed needed.

One group of learners that often strives for a nativelike accent is future teachers of the language at hand. Being a language teacher is one of the professions in which the pros and cons of being a native versus nonnative speaker of the language being taught have been discussed both within academia and within the public realm. If language is what is being taught, then some of the relevant questions are: Does the teacher need to possess certain linguistic qualities in order to adequately teach the language? By whom are those certain linguistic qualities to be judged? And is the judgement prone to language discrimination (Lippi-Green, 2012), a real and pervasive problem in many professions where a high level of language use is key (Moussu & Llurda, 2008)?

Researchers (Kamhi-Stein, 2004; Moussu, 2010; Park, 2012) have reported that native-speaking (NS) teachers receive preferential treatment (e.g. higher rates of hire, higher salaries, more employment perks) over nonnative-speaking (NNS) teachers in workplaces where both can be (or are) hired. This has continued even after several studies in the 1990s showed that NNS teachers may have certain advantages over NS teachers, such as a greater command of adult language learning principles (Phillipson, 1992), shared notions of

what it means to (and how to) learn the target language (Kramsch, 1997), and (when they have the same L1 as the learners do) the ability to use the learners' first language (L1) to explain complex concepts and to serve as a role model (Reves & Medgyes, 1994). Researchers have shown that students learn equally with NS and NNS teachers (Jacobs & Friedman, 1988), and that some students prefer NNS teachers (Ferguson, 2005; Meadows & Muramatsu, 2007). English as a world language is expanding and is taught in locations where NS teachers are unavailable, so hiring NNS teachers in many cases is unavoidable (for a review of research on NS versus NNS language teachers, see Moussu & Llurda, 2008). Nonetheless, NNS teachers worry about how others will respond to their accents (Park, 2012) and rightfully, in some contexts, they worry if an accent will keep them from being hired.

We know from research that employers often believe that native speakers make for better language teachers. For example, approximately 60% of American English language programme administrators and 72% of British administrators indicated in surveys that the primary factor considered when hiring new teachers is nativeness (Clark & Paran, 2007; Moussu, 2010). We also know from policy implementations that politicians or the public at large often believe that native speakers make the best teachers. For example, from 2002 to 2010, in the US state of Arizona, the Arizona Department of Education monitored thousands of English language teachers for accentedness (see Hanna & Allen, 2012, for a review and commentary on the policy). They did this because of a 2002 law that required that all English language teachers be qualified to teach English. This was outlined in the US No Child Left Behind Act (specifically, Title III, Section 3116-c), and Arizona interpreted teacher qualification in terms of accent. State accent monitors advised 'nonfluent' pre-K through 12th grade teachers to take accent reduction courses (Strauss, 2010). The policy ended in 2010 after the US Departments of Justice and Education launched an investigation into Arizona's policy: it was criticized for violating the teachers' civil rights. The eight-year policy highlights the debate over whether NS teachers are more effective than NNS teachers due to their accents (Kossan, 2011). The Arizona policy also stimulated, on a national level, a debate over why one would evaluate language teachers' accentedness, and how (or whether) pronunciation should be part of that evaluation. The *New York Times* (Lacey, 2011) questioned the motives of those mandating accent or pronunciation assessments under Arizona's policy. Who was actually qualified to do the evaluations, and why should they do this? Does accent and pronunciation assessment (and mandatory accent reduction) aid language teaching or society as a whole?

In our state of Michigan, in order to be certified to teach a language in the public schools, teachers must pass an Oral Proficiency Interview (OPI) exam which assesses whether the teachers' 'speech can be understood by native speakers unaccustomed to dealing with nonnatives' (as outlined in the American Council on the Teaching of Foreign Languages Advanced-Low

speech descriptor; see ACTFL, 2012: 11). This is because, by state law, teachers of the major foreign languages (Spanish, French, German) must be at the ACTFL Advanced-Low level at the time of certification, regardless of the teachers' teaching skills. Teacher candidates in Michigan can demonstrate this ability by passing the ACTFL OPI or another state-recognized proficiency exam, but it is up to the raters of those exams to qualify *understandable* speech. One ACTFL descriptor of speech at the Advanced-Low level is that the speakers can 'convey their intended message without misrepresentation or confusion' (ACTFL, 2012: 11), but again individual raters, even though they are trained, may perceive the comprehensibility of speech differently based on their familiarity with the speaker's L1 (as demonstrated by Gass & Varonis, 1984; Harding, 2012; Winke *et al.*, 2013; see Browne & Fulcher, this volume) or their notions or conceptions of the value of the speaker's accent (Gluszek & Dovidio, 2010; Lindemann, this volume). Why are raters or speech evaluators the gatekeepers? What about the students? What does accent mean to students? This part of the equation has been missing from the literature. It is this aspect of the teacher acceptability equation that we want to explore. Do the students themselves believe that NS teachers are better at teaching English? We question whether students evaluate or notice their teachers' accents or pronunciation, whether they divide their teachers into NS and NNS teacher categories, and whether teachers' accents or pronunciation matter to the students. We wonder this because students are the largest stakeholders within the realm of language teaching. Yet little research has been conducted on whether or how students perceive their language teachers' accents, and how their accents are related to their ability to be understood. Before we delve into our specific research questions, we first review the research on students' attitudes towards accents and accent perception.

Accent perception and students' attitudes towards accents

In assessing listeners' attitudes towards speakers with differing accents, one question that has been researched is whether listeners are able to accurately detect nativeness and identify accent. Derwing and Munro (2009: 476) defined an accent as 'either dialectal differences attributable to region or class, or phonological variations resulting from L1 influence on the [second language] L2', but they also described accent as 'the ways in which their [immigrants'] speech differs from that local [language] variety'. Thus, accent can be viewed from two broad theoretical perspectives: as various types of speech patterning that all individuals possess when speaking a language (hence, all language is accented); or on a societal level, as non-standard speech patterns spoken by individuals who are not native to the targeted language area, be they *foreigners* (possessing a foreign accent influenced by a different L1), or from a different geographical *region* (with the same L1, but

possessing a regional accent). Identifying an accent equates to being able to identify from where (which region, be it a nearby region or a foreign one) or from which L1 background the speaker stems.

Most research has shown that NSs are successful at identifying the nativeness and accent of NNSs (Derwing & Munro, 1997; Munro *et al.*, 2010). Whether listening to paragraphs or just syllables from a speaker, NSs are able to detect a nonnative accent in their L1 because they can rapidly detect a deviation from the standard native accent held in their mind (Flege, 1984), for example, being cued by prosodic dimensions of pronunciation such as rhythm (see Galaczi *et al.*, this volume). On the other hand, learners of a language do not easily distinguish among different regional or foreign accents of their second language, whether native or nonnative. Moussu (2010), who investigated the effects of accent exposure over time on students' attitudes towards their teachers, found that some students misidentified the native status of their teacher (i.e. classifying someone as a NS when they were a NNS, or vice versa), with whom they had studied for an entire semester. In Scales *et al.*'s (2006) study on accent perception, learners listened to four accents of the language they were learning (two NSs, two NNSs) and had difficulty in distinguishing which were NSs and NSSs. These studies suggest that, for learners, teacher accents may not matter because the learner may not readily differentiate this type of defined speech quality.

Apart from measurable perceptual differences present in regional or foreign accents, perceptual dialectologists have shown that accents also carry social stereotypes (see Lindemann, this volume; Preston, 1999). In the US context, Alford and Strother (1990) investigated NNSs' attitudes towards Northern, Southern and Midwestern US accents through subjective characteristic ratings. In general, the authors uncovered evidence of stigmas related to accents, that is, the notion that certain accents in certain regions or groups were associated with negative stereotypes (see Gluszek & Dovidio, 2010, for a review; and Robb, 2014, for more on accents and stereotypes). Importantly, as Alford and Strother (1990) found, NNSs attach these stigmas to certain dialects, too. Unlike previous studies, this study provides counterevidence that NNSs can differentiate between accents.

Students' attitudes towards their teachers' accents

Within foreign language pedagogy, researchers have investigated students' perceptions of their language teachers' accents. Such research has found that students' attitudes do not hinge only on their teachers' native status. Teachers' L1s are one of many variables that affect how students view their teachers, other factors being students' expected grades, students' majors, and teachers' country of origin (Moussu, 2010). In Moussu's study, students' impressions of their NNS teachers became more positive over the semester. While students overtly indicated that they preferred NS teachers,

an implicit association test suggested that students, in fact, valued NNS teachers equally. Hertel and Sunderman (2009) found that, while students reported that they preferred to learn pronunciation from NS teachers, they appreciated NNS teachers' abilities to give grammatical rules and explain vocabulary, and they also benefited from exposure to NNS teachers and their accents, resulting in increased NNS accent comprehension (Gass & Varonis, 1984; Winke *et al.*, 2013). Additionally, when students share the same L1 as the teacher, NNS teachers may use the shared L1 to facilitate comprehension. Students themselves report better understanding of their teacher when the L1 is shared because the accent is easier to understand (Hertel & Sunderman, 2009; Park, 2012). These studies provide a snapshot of students' general positive attitudes towards their NNS teachers.

Student attitudes aside, many researchers have examined aspects of the perceptual processing of accents. Findings have shown that *familiarity*, operationalized by Derwing and Munro (1997) as the amount of contact one has had with a particular accent, has a positive impact on listeners' comprehension. In fact, research suggests that language learners' background knowledge and linguistic experiences contribute to comprehension more than teachers' accents detract from it (Derwing & Munro, 1997; Gass & Varonis, 1984; Harding, 2012; Winke *et al.*, 2013). When listeners are familiar with a speaker's speech variety, they display better listening comprehension despite the speakers' accents (Gass & Varonis, 1984). Moreover, when language learners are familiar with a speaker's L1 because it is the same as their own L1 (the shared L1 advantage; Harding, 2012, this volume), or because they have grown accustomed to the L1 through sufficient exposure (Gass & Varonis, 1984; Winke *et al.*, 2013), they display better comprehension of the L2 speech of speakers with those L1s. Bradlow and Bent (2008) found that individuals are highly flexible in processing foreign-accented speech. Individuals learn the patterns and then apply the pattern knowledge when they come across (new) people who speak with the same accent. These findings suggest that it is only a matter of time before students become comfortable with new NNS accents.

What is particularly important is whether learners can understand their teachers (find them comprehensible and intelligible). Understanding accented speech is split along *comprehensibility* and *intelligibility* dimensions because a strong accent does not necessarily impede intelligibility, but any type of accent may take more effort or time to process, to render it comprehensible (see Derwing & Munro, 1997). Speech is comprehensible if it is easy to understand (in comparison to incomprehensible speech that this difficult or impossible to understand); this is a judgement call in relation to the internal, cognitive effort it takes for a listener to process speech. On the other hand, speech intelligibility, or the amount of speech one can understand, can be measured more objectively, through the accuracy of listeners' orthographic transcriptions, for instance. Students may be concerned about the

comprehensibility and the intelligibility of their teachers. (See Saito *et al.*, this volume, for a discussion of the complexity of comprehensibility.)

The Current Study

Thinking back to the accent legislation in Arizona, there was no evidence concerning how students, arguably the most important stakeholders in this issue, felt about the quality of their education and the acceptability of their teachers with regard to their teachers' accent. If they found their teachers, regardless of their accent, to be intelligible and comprehensible, why would accent matter at all? The following questions guided our study:

(1a) Are NNS students able to distinguish NSs from NNSs of English?
(1b) Can NNS students identify the speakers' accents?
(2a) When students are familiar with an accent, do they rate the speakers more favourably in terms of comprehensibility, intelligibility, accentedness, and acceptability as a teacher?
(2b) Is there a relationship between students' comprehensibility, intelligibility, and accentedness ratings and their attitudes about a speaker's acceptability as an English teacher?
(3) Does nativeness account for any variance in ratings of acceptability as an English teacher after comprehensibility, intelligibility, and accentedness are accounted for?

Methodology

Participants

We recruited 121 participants from eight classes at a large Midwestern university (M_{age} = 21.7, range = 18–51). The participants included 85 international students, the majority of whom (87%) were enrolled in part-time or full-time ESL classes at three proficiency levels: Intermediate-Low (26), Intermediate-High (25) and Advanced-Low (30) on the ACTFL scale. The remaining NNS students were undergraduate (1), Master's (1) and doctoral students (1). The students spoke Chinese (n = 52; 61%), Arabic (n = 19; 22%), Korean (n = 6; 8%) or a different language (n = 8; 9%). We also recruited a comparison group of 36 American English speakers from two undergraduate-level courses.

Materials

We used SurveyMonkey.com, an online survey system, to create a web-based survey. All data were collected using this online survey.

Background questionnaire

This asked about students' language learning and their exposure to specific accents. Most important for this study, the students' familiarity with the accent of each speech sample presented was measured on separate five-point Likert-type scales ranging from 1 (very familiar) to 5 (not familiar at all).

Listening tasks

Using professional equipment at a digital recording studio, we recorded three NSs (American Midwestern from Michigan, British from northern England, and American Southern from Alabama) and two NNSs (Chinese, Albanian) speech samples. Each speaker read and recorded: (a) a paragraph about a familiar topic (ESL classroom expectations; 20–24 seconds); and (b) a paragraph about an unfamiliar topic (pottery making, adapted from Sueyoshi & Hardison, 2005; 20–24 seconds).

Likert-scale items

The nine-point Likert scale items elicited responses about the following variables:

(1) *Intelligibility*: the listener's estimation of his or her understanding of a speaker's utterance, with the student being asked, 'How much of this speech did you understand?' as measured through a rated degree of understanding (Derwing & Munro, 1997, 2009) on a nine-point scale ranging from '100%, everything' (1 point) to 0%, nothing (9 points).
(2) *Comprehensibility*: 'the listener's estimation of difficulty in understanding an utterance' (Munro *et al.*, 2010: 112), measured on a nine-point scale ranging from 'very easy to understand' (1 point) to 'very difficult to understand' (9 points).
(3) *Accentedness*: 'the degree to which the pronunciation of an utterance sounds different from an expected production pattern' (Munro *et al.*, 2010: 112), measured on a nine-point scale ranging from 'no accent' (1 point) to 'very strong accent' (9 points).
(4) *Acceptability as a teacher*: the listener's estimation of how acceptable the speaker is as an ESL teacher, measured on a nine-point scale ranging from 'acceptable' (1 point) to 'not acceptable' (9 points).

These rating scales were used for each audio clip that the students rated, with two speaking task performances from each speaker.

Nativeness and accent

After listening to each speaker's recording, students indicated whether they thought speakers were NSs (yes or no). They then indicated what they thought the speakers' accents were from a given list, which included the

target accents and plausible distractors (I don't know, Albanian, American Midwestern, American Southern, Australian, British, Chinese, French, Indian, Japanese, Malagasy, Nigerian, Spanish). For the audio-based tasks, the students could have listened to any audio file as many times as they wanted even though they were instructed to listen only once to each individual sound file. This could have yielded unequal levels of exposure, and we recognize that this is a limitation to our study.

Procedure

Students either met in a computer lab in groups (110) or completed the web-based survey on their own time (11). All students first completed a familiarization task similar to the target listening tasks where they listened to one speaking performance and then responded to nine-point Likert scale questions. Next, students completed the background questionnaire and the experimental listening tasks. To eliminate ordering effects, the speech samples were presented randomly. The students listened to each speaking performance twice, answering two Likert scale questions after each time (four total items per speech sample and eight total items per speaker, for five speakers, yielding 40 questions per student). Afterwards, they evaluated the speaker's NS/NNS status and accent.

Results

Research Question 1

In relation to Research Question 1a (Are NNS students able to distinguish NSs from NNSs?), we found that NNS students correctly identified the speakers' native status 68% of the time, while NS students did so 91% of the time. NNS students were significantly less able to identify the accents of all speakers in the study except for that of the British speaker, for whom neither group was particularly successful. An independent samples *t*-test comparing the group means indicated that NNS students were significantly less able than NS students to distinguish the speakers' nativeness, $t(102) = -8.91$, $p < 0.001$, *Cohen's d* (effect size) = 1.65 (see Table 7.1).

In relation to Question 1b (Can NNS students identify the speakers' accents?), we found NSs correctly identified the speakers' accent 57% of the time, while NNSs did so 26% of the time (Table 7.2). The NS group outperformed the NNSs in every identification except for Chinese. An independent samples *t*-test comparing group means of correct identification revealed that the NNS students were significantly less able than the NS students to distinguish the speakers' accents, $t(80) = -9.88$, $p < 0.001$, $d = 1.87$ (see Table 7.2).

Table 7.1 Group means (*SD*) for correct identification of speaker status as NS or NNS

Speaker	*NSs*	*NNSs*
Albanian	67 (24) = 93%	79 (36) = 46%
British	45 (44) = 63%	92 (42) = 54%
Chinese	71 (8) = 99%	138 (30) = 81%
Midwestern	72 (0) = 100%	106 (36) = 62%
Southern	71 (8) = 99%	128 (29) = 75%
M	65 (10) = 91%	116 (17) = 68%

Notes: Number of correct identifications (percentage correct) for NS ($n = 36$) and NNS ($n = 85$) participants, with each participant making two identifications per speaker. The overall difference between NS and NNS participants was significant, $t(102) = -8.91$, $p < 0.001$, $d = 1.65$.

Table 7.2 Group means of correct accent identification and *t*-test results

Speaker	*NSs*	*NNSs*
Albanian	12 (38) = 17%	5 (17) = 03%
British	42 (50) = 58%	50 (46) = 29%
Chinese	28 (49) = 39%	74 (50) = 44%
Midwestern	68 (23) = 94%	56 (47) = 33%
Southern	54 (44) = 75%	33 (40) = 19%
M	41 (15) = 57%	44 (18) = 26%

Notes: Number of correct identifications (percentage correct) for NS ($n = 36$) and NNS ($n = 85$) participants, with each participant making two identifications per speaker. The overall difference between NS and NNS participants was significant, $t(80) = -9.88$, $p < 0.001$, $d = 1.87$.

To explore whether proficiency affected NNS students' abilities to identify speakers' accents, we compared accuracy ratings across the different levels of proficiency. We excluded NNS students who did not have specific proficiency information available ($n = 3$). The data show a clear trend: when the proficiency level increases, the ability to identify NS/NNS status and accent increases (Table 7.3). The exception to this trend is between the Intermediate-High and Advanced-Low group, in which there were small or no gains.

Research Question 2

To answer Question 2a, we first examined composite ratings of familiarity with the five accents. On a scale of 1 ('not familiar') to 5 ('very familiar'), NNSs reported a familiarity of 1.33 ($SD = 0.7$) with Albanian, 2.93 ($SD = 1.38$) with British, 3.6 ($SD = 1.46$) with Chinese, 3.41 ($SD = 1.27$) with

Table 7.3 Correct identification of native/nonnative status and L1 accent by NNS students by proficiency level

NNSs' proficiency	*n*	*Native status*	*L1 accent*
Intermediate low	26	149 (57%)	48 (18%)
Intermediate high	25	166 (66%)	64 (26%)
Advanced low	30	196 (65%)	87 (29%)
Total	81		

Notes: Total correct identifications (percentage correct); each participant made two identifications for each of five speakers. Proficiency levels are on the ACTFL Proficiency Scale (http://www.actfl.org).

Midwestern and 2.99 ($SD = 1.27$) with Southern accents. The NSs reported a familiarity of 1.64 ($SD = 1.15$) with Albanian, 3.78 ($SD = 1.1$) with British, 3.06 ($SD = 1.15$) with Chinese, 4.92 ($SD = 0.37$) with Midwestern and 4.67 ($SD = 0.48$) with Southern accents. The familiarity ratings are visually represented in Figure 7.1.

Next, we calculated Spearman's rank order correlations to uncover relationships between NNSs' reported familiarity with an L1 and their ratings of speakers' comprehensibility, intelligibility, accentedness, and acceptability as teachers (Table 7.4). Students' self-reported familiarity with an accent was significantly correlated with their ratings of all four variables for Chinese (acceptability: $r_s = -0.29$, $p < 0.001$; accentedness: $r_s = -0.37$, $p < 0.001$; intelligibility: $r_s = -0.41$, $p < 0.001$; comprehensibility: $r_s = -0.38$, $p < 0.001$), and three variables for British English (acceptability: $r_s = -0.25$, $p < 0.05$; intelligibility: $r_s = -0.36$, $p < 0.001$; comprehensibility: $r_s = -0.25$, $p < 0.05$). All other coefficients were nonsignificant.

To investigate whether students' ratings were related to their attitudes about acceptability as an English teacher, we calculated Spearman's correlations (Table 7.5). Results indicated that acceptability as a teacher statistically correlated with accentedness, intelligibility and comprehensibility for the ratings of every speaker (Albanian, British, Chinese, Southern and Midwestern). A telling

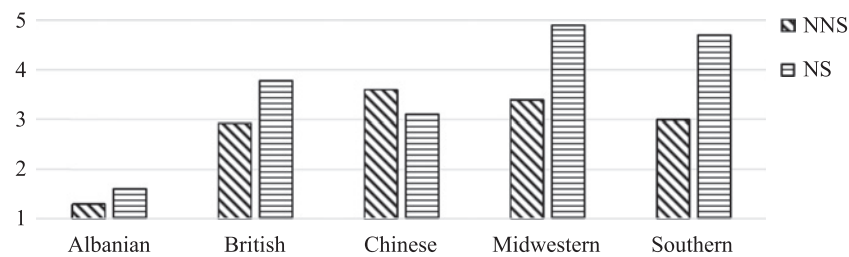


Figure 7.1 Composite ratings of accent familiarity by all students

Table 7.4 Spearman's correlations for NNSs' reported accent familiarity and students' ratings of four variables for corresponding speakers

Speaker	*n*	*Acceptability*	*Accentedness*	*Intelligibility*	*Comprehensibility*
Albanian	82	–0.05	0.01	0.09	0.01
British	83	–0.25*	0.06	–0.36**	–0.25*
Chinese	84	–0.29**	–0.37**	–0.41**	–0.38**
Midwestern	83	–0.17	–0.02	–0.08	–0.11
Southern	85	–0.02	–0.01	–0.03	–0.08

Note: *$p < 0.05$, **$p < 0.01$, two-tailed.

Table 7.5 Spearman's correlations between NNSs' ratings on teachers' acceptability and teachers' accentedness, intelligibility and comprehensibility

Speaker	*n*	*Accentedness*	*Intelligibility*	*Comprehensibility*
Albanian	82	0.44**	0.64**	0.71**
British	83	0.33**	0.71**	0.80**
Chinese	84	0.49**	0.57**	0.69**
Midwestern	83	0.48**	0.76**	0.64**
Southern	85	0.50**	0.71**	0.71**

Note: *$p < 0.05$, **$p < 0.01$, two-tailed.

trend that emerged in these data is that acceptability was least strongly correlated with accentedness, which showed moderate correlations ($r_s = 0.33–0.50$), followed by intelligibility ($r_s = 0.57 – 0.76$) and comprehensibility ($r_s = 0.64 – 0.80$), which both show moderate to strong correlations with acceptability. Acceptability ratings are plotted visually in Figure 7.2.

Research Question 3

We performed a multivariate regression to see if nativeness contributed to the regression model after taking comprehensibility, intelligibility and accentedness into account. To run the regression, we averaged the two scores that each listener gave each speaker for each task in order to have one value per speaker. Then we replaced any data for outliers with the mean plus three standard deviations; this was done in order to keep, rather than exclude, data points when considering the best model fit. Next, we did a log transformation on the values for accentedness in order to have the best linear fit possible ($R^2 = 0.23$). After making these modifications, the data satisfied the assumptions for a multiple regression (see Field, 2009). To account for repeated measures within the data, we performed a

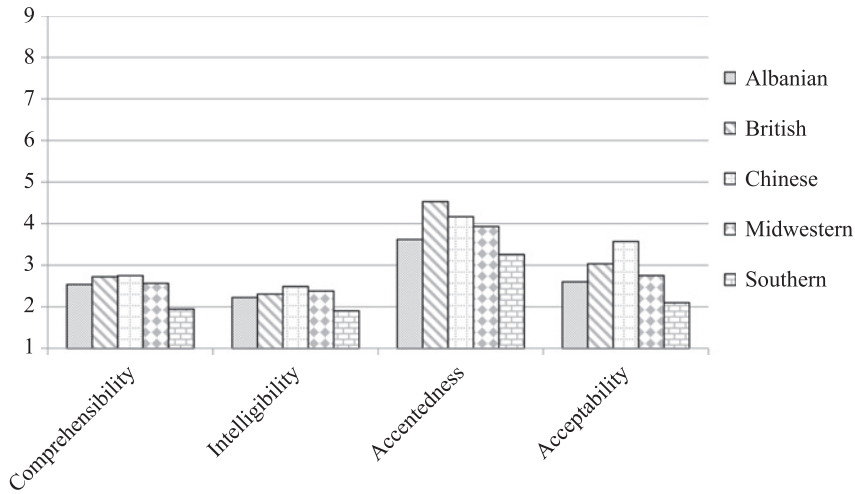


Figure 7.2 Composite ratings of speakers on four variables by all NNS students
Notes: 1 = 'acceptable as an English teacher/no accent', 9 = 'not acceptable as an English teacher/strong accent'. The highest average score on acceptability was 4.50, so only the bottom end of the scale is displayed.

multivariate linear regression with a random subject factor. The results of the regression are presented in Table 7.6. Adding nativeness to the model showed that this factor did, in fact, account for variation in acceptability scores. Looking at the regression gradients, however, is telling of the impact that perceived nativeness had on acceptability scores. The more native the students perceived the speakers, the more acceptable as teachers the students found the speakers.

Table 7.6 Multivariate linear regression parameter estimates

Parameter	*B*	*SE*	*95% Wald CI*		*Hypothesis test*		
			Lower	*Upper*	*Wald* χ^2	*df*	*p*
(Intercept)	−0.44	0.20	−0.83	−0.06	5.16	1	0.02
Comprehensibility	0.41	0.10	0.21	0.61	15.53	1	<0.01
Intelligibility	0.41	0.12	0.18	0.65	12.17	1	<0.01
Accentedness	0.94	0.30	0.36	1.53	9.93	1	<0.01
Native guess	0.82	0.18	0.47	1.18	21.15	1	<0.01
Split guess	0.44	0.15	0.14	0.75	8.23	1	<0.01

Notes: Dependent variable = acceptability. Model = (Intercept), comprehensibility, intelligibility, accentedness, native guess, split guess. *SE* = standard error; CI = confidence interval.

'Nativeness guess' indicates that, if a student thought a speaker was a native English speaker, his or her rating for acceptability improved by 0.80 of a point (towards acceptable and away from unacceptable), such as from 2.50 to 1.70. In other words, listeners rated individuals they identified as native English speakers as more acceptable. For 'split guess' (which means that for one speaking performance a student thought the speaker was a native English speaker, and for the other nonnative), the rating for acceptability improved by 0.40 points. This also showed that listeners rated a speaker as more acceptable English teachers if they thought the speaker was native in one speaking performance and nonnative in the other.

Discussion

We aimed to gain a greater understanding of student attitudes towards NS and NNS teachers by assessing students' abilities to perceive accents and their reactions to different native and nonnative English accents. We did this because language teachers, regardless of their overall teacher qualifications, are sometimes assessed on pronunciation, but pronunciation is often operationalized in terms of foreign accent, as it was in Arizona from 2002 to 2010. And this accent evaluation is done irrespective of students' abilities to understand (and learn from) the teachers.

Language teacher certification processes in the United States, and indeed language teaching hiring processes around the world, often involve assessments (formal or informal) of the teachers' oral proficiency. Accent bias that surfaces during such assessments may lead towards the hiring of NS teachers over NNS teachers (Gluszek & Dovidio, 2010), even though the NNS teachers may have better teaching skills (as summarized by Hertel & Sunderman, 2009; Moussu, 2010). Or, as described in the literature, NS teachers may be hired first and foremost over NNS teachers (the so-called *native speaker fallacy*, as observed by Clark & Paran, 2007; Kahmi-Stein, 2004; Moussu, 2010; Park, 2012) because NSs are believed to be best at teaching, mainly because they are considered to have the language variety learners may want (Scales *et al.*, 2006) and because they are able to serve as expert cultural ambassadors.

In this study, we have extended previous research on teacher accentedness to encompass student attitudes. Our results suggest that NNS students were often able to distinguish between NSs and NNSs. However, they were generally unable to identify a speaker's accent. Results further suggest that familiarity with an accent is positively correlated with comprehensibility and acceptability as a teacher. Additionally, students had generally positive attitudes towards nonnative accents. We suggest those students' attitudes towards their NNS teachers would only become more positive over time, as students will rapidly perceptually adapt to their teachers' accents. Finally,

this research indicated that, in the minds of students, accentedness does not translate to unacceptability as a teacher.

Speaker identification: Status and accent

In looking at students' abilities to identify a speaker's native status and their accent, the results from this study corroborate those of Derwing and Munro (1997). NSs outperformed NNSs in these tasks. However, Derwing and Munro reported that NNSs were much more accurate in identifying accents (52%) than in the present study (25%). In their study, participants selected an accent from a small pool of options (four), while in our study students had to distinguish accents from a much larger option pool (12). This larger range of choices are likely to have contributed to higher NNS inaccuracy, although it is arguably closer to real-life circumstances in which listeners must distinguish between numerous accent options.

We believe familiarity also partially explains the NNS accent identification inaccuracy shown in this study. It appears that students misidentified Albanian accents and American Southern accents because they were unfamiliar with both (as shown from the familiarity ratings in their background questionnaires). The majority of NNS students had no exposure to Albanian and very little exposure to an American Southern accent because of the geographic location of this study (in the upper Midwest) and a general lack of representation of these accents in the media. Whereas identification of the Albanian and Southern accents varied in the data, identification of Midwestern, British and Chinese accents was more accurate and less varied. This consistency could be attributed to the learners' greater knowledge and familiarity with these cultures and accents, as reported in the background questionnaires. From the data, we conclude that familiarity played a positive role in their ability to identify accents.

Another possible explanation for the NNSs' low identification accuracy is that they did not know how the native accents really sound. An inaccurate concept of what a NS is supposed to sound like could lead to inaccurate judgements (Scales *et al.*, 2006), which could explain why only 30% of the NNS in our study were able to identify the native-English speakers (Midwestern, Southern and British).

Proficiency level also played a role in the students' abilities to identify accents. Our data show that as proficiency increased, the ability to distinguish native and nonnative accents and the ability to identify specific accents also increased (Table 7.4). Similar results have been described before. Beinhoff (2014) investigated how learners of English at different proficiency levels perceived the accentedness and intelligibility of Spanish-accented English. She found that, although receptive skills are not described in language proficiency scales as part of phonological control (which is mainly described in terms of production), perception and production do interact. She referenced

Flege's Speech Learning Model (1995) to explain why higher proficiency learners had fewer problems with the Spanish-accented speech samples' intelligibility than lower level learners did. Part of Flege's model suggests that accurately perceiving new sounds in the target language is a necessary precursor to being able to produce the sounds. As learners' proficiency level increases, the amount and variation of new sounds acquired increases, making slight inter-speaker variations in vowel length and quality, variations in consonants, and insertions of vowels that break up consonant clusters less influential on the overall intelligibility and comprehensibility of the L2 (see Beinhoff, 2014: 67). We expect that the same happened here.

Although the NNS students did have trouble making correct accent identifications, as Lindemann (2003) suggested, even if a listener is unable to correctly identify an accent, positive or negative stereotypes associated with that accent can be activated and can, therefore, influence a listener's attitudes towards the speaker. Thus, we are confident that these students' belief ratings are informative despite their inability to label specific accents.

The role of accent familiarity

In investigating the relationship of accent familiarity with comprehensibility, intelligibility, accentedness, and acceptability as a teacher, the results showed weak correlations, indicating that familiarity with an accent is significantly related to students' judgements about: (a) how easy it is to understand a speaker of that accent; and (b) the speaker's acceptability as a teacher (Table 7.5). The more understandable the accent, the more acceptable the speaker was perceived as a teacher. This significant relationship between familiarity and comprehensibility aligns with findings described by Gass and Varonis (1984) and Winke *et al.* (2013), who showed that accent familiarity facilitates comprehension.

This result gave us great pause. In the New York Times article (Lacey, 2011), many of the English language teachers who failed the accent evaluation assessment in Arizona (and who were sent to accent reduction courses) were described as NSs of Spanish, which is the L1 of most of the students in the English language classes in Arizona. The speech evaluators, however, tended to *not* be NSs of Spanish, but rather were NSs of English with varying levels of familiarity with Spanish-accented speech. So the pronunciation and accent evaluations in Arizona may have resulted in reducing the number of teachers in the classroom who shared an L1 with the students, something that has been shown to be beneficial in language learning programmes (Hertel & Sunderman, 2009; Marian *et al.*, 2008; Park, 2012). As in Scales *et al.*'s (2006) and Isaacs' (2008) research, we found that comprehension was a high priority for students in accepting a speaker as a teacher. However, teacher acceptability based on accent or pronunciation alone may not be

appropriate, especially when that acceptability judgement may be tied to negative stereotypes (Gluszek & Dovidio, 2010).

In Moussu's (2010) study, students' attitudes towards their NNS teachers became more positive over the semester. Thus, we would like to caution that students' (or anyone's) initial judgements of a teacher's acceptability (based on comprehensibility and intelligibility) should be considered with care. Research has shown that NSs can rapidly adapt to accented speech through exposure (Bradlow & Bent, 2008; Clarke & Garrett, 2004). And even if actual understanding does not improve (and it might not if listeners' proficiency does not increase), research does suggest that NNSs might, over time, gain confidence about dealing with other NNSs (Derwing *et al.*, 2002), and that extra confidence may maximize comprehension. Thus, even if students are wary of a teacher's accent on first exposure, those in charge should understand that this hesitation may lessen over time and through exposure. Moreover, the time and exposure to the accent may benefit the students: prolonged exposure to different accents helps learners better comprehend the language in general (Clarke & Garrett, 2004), and this is more reflective of the real-world spoken language that learners will encounter along their language learning journey (see Wagner & Toth, this volume, for a discussion of the benefits of authentic and natural listening contexts).

Student perceptions and native speaker demand

The final analysis examined whether nativeness accounted for any variance in acceptability after comprehensibility, intelligibility and accentedness were accounted for. Although the statistical addition of nativeness to the model shows that the students favoured NSs as English teachers, it is important to consider the regression gradients in light of the acceptability scale they were measured on and its standard deviation. Because acceptability was measured on a scale from 1 ('acceptable') to 9 ('not acceptable'), a movement of 0.70 of a point does not seem very impactful. This is important considering the fact that this rating (based on the mean of 2.40) still falls well within the standard deviation of the acceptability rating. To illustrate, if a student thought they listened to a nonnative speaker, they could have given a 2.40 (mean for acceptability). If the same student thought that the speaker was native, their rating was likely to be 1.70, with a split decision falling somewhere in between. Considering that the scale ranged from 1 to 9, this movement from 2.40 to 1.70 seems inconsequential. We argue that, whether the student thought the speaker was native in both speaking performances (acceptability = 1.70), was split on the performances (acceptability = 2.00), or thought the speaker was nonnative in both performances (acceptability = 2.40), the student effectively assigned similar scores for acceptability. This is further evidence, corroborating that of previous

studies, against native speaker fallacy: while students say that they prefer NS teachers, their attitudes don't strongly reflect that sentiment (Moussu, 2010; Phillipson, 1992).

Implications

Student acceptance of nonnative accents provides evidence that challenges the assumptions of language centre administrators, such as those in Moussu's (2010) and Clark and Paran's (2007) studies, who reported that one of the major factors in hiring English language teachers is NS status. Administrators claimed that they continue these hiring practices because of 'native speaker demand' on the part of students. Overall, our results suggest that student attitudes towards NNS teachers may be positive even when students indicated that a teacher had a pronounced accent, as the students still rated them as acceptable. This shows that, from a student's perspective, factors other than accent may more heavily influence a student's attitude towards a teacher (Kang, 2012). Furthermore, exposing language learners to a variety of accents has the potential to equip them with better listening skills (Clarke & Garrett, 2004; Gass & Varonis, 1984); thus the exposure would benefit students. This reality should be considered when hiring language teachers. Those hiring should not automatically pass nonnative teachers by; rather, they should consider that the teachers' accentedness may actually be an asset in helping students become better language learners.

Conclusion

As seen in the current study, the relationships among accentedness, accent perception ability, and student beliefs are complex. We have investigated the interplay of these concepts, but more research is needed to further clarify these relationships. In the future we see two lines of research being particularly helpful in doing so. First, in a classroom-based mixed-methods study, it would be valuable to track NNS students in classrooms with NNS teachers over the course of a semester, looking specifically at (a) how student beliefs about their teachers change in relation to (b) changes in their actual perceptual adaptation (i.e. increased comprehensibility and intelligibility) to their teachers' accents. Secondly, it would be valuable to empirically investigate the rate of NNS adaptation to NNS speech among speakers from different L1 backgrounds (see Clarke & Garrett, 2004, for a similar study with NSs). Information gleaned from these types of studies could further clarify the dynamic relationship between cognitive processes and social beliefs tied to NNS accents.

References

ACTFL (2012) Oral proficiency familiarization manual. White Plains, NY: American Council on The Teaching of Foreign Languages. See http://www.languagetesting.com/wp-content/uploads/2013/05/ACTFL-OPI-Familiarization-Manual1.pdf.

Alford, R.L. and Strother, J.B. (1990) Attitudes of native and nonnative speakers toward selected regional accents of U.S. English. *TESOL Quarterly* 24 (3), 479–495.

Beinhoff, B. (2014) What is 'acceptable'? The role of acceptability in English non-native speech. In M. Solly and E. Esch (eds), *Language Education and the Challenges of Globalisation: Sociolinguistic Issues* (pp. 155–174). Cambridge: Cambridge Scholars.

Bradlow, A.R. and Bent, T. (2008) Perceptual adaptation to non-native speech. *Cognition* 106 (2), 707–729.

Clark, E. and Paran, A. (2007) The employability of non-native-speaker teachers of EFL: A UK survey. *System* 35 (4), 407–430.

Clarke, C.M. and Garrett, M.F. (2004) Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America* 116 (6), 3647–3658.

Crowther, D., Trofimovich, P., Isaacs, T. and Saito, K. (2015) Does a speaking task affect second language comprehensibility? *Modern Language Journal* 99 (1), 80–95.

Derwing, T.M. and Munro, M.J. (1997) Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition* 19 (1), 1–16.

Derwing, T.M. and Munro, M.J. (2009) Putting accent in its place: Rethinking obstacles to communication. *Language Teaching* 4 (4), 476–490.

Derwing, T.M., Rossiter, M.J. and Munro, M.J. (2002) Teaching native speakers to listen to foreign-accented speech. *Journal of Multilingual and Multicultural Development* 23 (4), 245–259.

Ferguson, A. (2005) Student beliefs about their foreign language instructors: A look at the native-speaker/non-native speaker issue. *Dissertation Abstracts International, A: The Humanities and Social Sciences* 522-A–523-A.

Field, A. (2009) *Discovering Statistics Using SPSS* (3rd edn). London: Sage.

Flege, J.E. (1984) The detection of French accent by American listeners. *Journal of the Acoustical Society of America* 76 (3), 692–707.

Gass, S. and Varonis, E.M. (1984) The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning* 34 (1), 65–89.

Gluszek, A. and Dovidio, J.F. (2010) The way they speak: A social psychological perspective on the stigma of nonnative accents in communication. *Personality and Social Psychology Review* 14 (2), 214–237.

Hanna, P.L. and Allen, A. (2012) Educator assessment: Accent as a measure of fluency in Arizona. *Education Policy* 27 (4), 711–738.

Harding, L. (2012) Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing* 29 (2), 163–180.

Hertel, T.J. and Sunderman, G. (2009) Student attitudes toward native and non-native language instructors. *Foreign Language Annals* 42 (3), 468–482.

Isaacs, T. (2008) Towards defining a valid assessment criterion of pronunciation proficiency in non-native English speaking graduate students. *Canadian Modern Language Review* 64 (4), 555–580.

Jacobs, L.C. and Friedman, C.B. (1988) Student achievement under foreign teaching associates compared with native teaching associates. *Journal of Higher Education* 59 (5), 551.

Kamhi-Stein, L. (2004) *Learning and Teaching from Experience*. Ann Arbor, MI: University of Michigan Press.

Kang, O. (2012) Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly* 9 (3), 249–269.

Kossan, P. (2011) Arizona teacher accent scrutiny halted. *The Arizona Republic*, 12 September. See http://archive.azcentral.com/arizonarepublic/news/articles/20110912arizona-teacher-accent-scrutiny-halted.html.

Kramsch, C.J. (1997) Culture and constructs: Communicating attitudes and values in the foreign language classroom. In P. Heusinkveld (ed.) *Pathways to Culture: Readings on Teaching Culture in the Foreign Language Class* (pp. 461–485). Yarmouth, ME: Intercultural Press.

Lacey, M. (2011) In Arizona, complaints that an accent can hinder a teacher's career. *The New York Times*, 25 September. See http://www.nytimes.com/2011/09/25/us/in-arizona-complaints-that-an-accent-can-hinder-a-teachers-career.html.

Levis, J.M. (2005) Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly* 39 (3), 369–377.

Lindemann, S. (2003) Koreans, Chinese, or Indians? Attitudes and ideologies about non-native English speakers in the United States. *Journal of Sociolinguistics* 7 (3), 348–364.

Lippi-Green, R. (2012) *English With an Accent: Language, Ideology, and Discrimination in the United States* (2nd edn). New York: Routledge.

Marian, V., Blumenfeld, H.K. and Boukrina, O.V. (2008) Sensitivity to phonological similarity within and across languages. *Journal of Psycholinguistic Research* 37 (3), 141–170.

Meadows, B. and Muramatsu, Y. (2007) Native speaker or non-native speaker teacher? A report of student preferences in four different foreign language classrooms. *Arizona Working Papers in Second Language Acquisition and Teaching* 14, 95–109.

Moussu, L. (2010) Influence of teacher-contact time and other variables on ESL students' attitudes towards native- and nonnative-English-speaking teachers. *TESOL Quarterly* 44 (4), 746–768.

Moussu, L. and Llurda, E. (2008) Non-native English-speaking English language teachers: History and research. *Language Teaching* 41 (3), 315–348.

Munro, M.J., Derwing, T.M. and Burgess, C.S. (2010) Detection of nonnative speaker status from content-masked speech. *Speech Communication* 52 (7–8), 626–637.

Park, G. (2012) 'I am never afraid of being recognized as an NNES': One teacher's journey in claiming and embracing her nonnative-speaker identity. *TESOL Quarterly* 46 (1), 127–151.

Phillipson, R. (1992) *Linguistic Imperialism*. Oxford: Oxford University Press.

Preston, D. (1999) *Handbook of Perceptual Dialectology*. Amsterdam: John Benjamins.

Reves, T. and Medgyes, P. (1994) The non-native English speaking EFL/ESL teacher's self-image: An international survey. *System* 22 (3), 353–367.

Robb, A. (2014) A person's accent can change your perception of what he is saying. *New Republic*, 23 September. See http://www.newrepublic.com/article/119546/accents-can-influence-perception.

Scales, J., Wennerstrom, A., Richard, D. and Wu, S.H. (2006) Language learners' perceptions of accent. *TESOL Quarterly* 40 (4), 715–738.

Strauss, V. (2010) How Arizona is checking teachers' accents. *The Washington Post*, 27 May. See http://voices.washingtonpost.com.

Sueyoshi, A. and Hardison, D.M. (2005) The role of gestures and facial cues in second language listening comprehension. *Language Learning* 55 (4), 661–699.

Tokumoto, M. and Shibata, M. (2011) Asian varieties of English: Attitudes towards pronunciation. *World Englishes* 30 (3), 392–408.

US Department of Education (2002) Public Law 107–110 – JAN. 8, 2002. See http://www.ed.gov/policy/elsec/leg/esea02/index.html.

Winke, P., Gass, S. and Myford, C. (2013) Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing* 30 (2), 231–252.

8 Re-examining Phonological and Lexical Correlates of Second Language Comprehensibility: The Role of Rater Experience

Kazuya Saito, Pavel Trofimovich, Talia Isaacs and Stuart Webb

Introduction

Few researchers and teachers would disagree that some linguistic aspects of second language (L2) speech are more crucial than others for successful communication. Underlying this idea is the assumption that communicative success can be broadly defined in terms of speakers' ability to convey the intended meaning to the interlocutor, which is frequently captured through a listener-based rating of comprehensibility or ease of understanding (e.g. Derwing & Munro, 2009; Levis, 2005). Previous research has shown that communicative success – for example, as defined through comprehensible L2 speech – depends on several linguistic dimensions of L2 output, including its segmental and suprasegmental pronunciation, fluency-based characteristics, lexical and grammatical content, as well as discourse structure (e.g. Field, 2005; Hahn, 2004; Kang *et al.*, 2010; Trofimovich & Isaacs, 2012). Our chief objective in the current study was to explore the L2 comprehensibility construct from a language assessment perspective (e.g. Isaacs & Thomson, 2013), by targeting rater experience as a possible source of variance influencing the degree to which raters use various characteristics of speech in judging L2 comprehensibility. In keeping with this objective, we asked the following question: What is the extent to which linguistic aspects of L2 speech contributing to comprehensibility ratings depend on raters' experience?

Linguistic correlates of L2 comprehensibility

The relationship between L2 comprehensibility and the linguistic content of L2 speech (e.g. in terms of its segmental, suprasegmental or fluency-based characteristics) has been a productive area of research. For instance, L2 comprehensibility appears to be related to various linguistic dimensions of L2 speech, including individual sounds with high functional load, such as those that distinguish word meaning across many word pairs (Munro & Derwing, 2006), and those that represent 'lingua franca core' sounds, such as vowels and consonants which frequently lead to miscommunication in interaction between L2 speakers (Jenkins, 2000). Beyond segmentals, understandable L2 speech also seems to be linked to the production of word stress (Field, 2005), sentence stress (Hahn, 2004), and such aspects of fluency as pausing frequency (Kang *et al.*, 2010).

Perhaps a question that is more relevant to both language researchers and teachers is not which linguistic dimensions of speech contribute to L2 comprehensibility, but rather which linguistic dimensions are relatively more important for comprehensibility, compared to other dimensions. For example, Derwing *et al.* (1998) showed that a 12-week course for ESL learners in Canada with an explicit focus on prosody (i.e. suprasegmentals, such as word stress) and fluency resulted in more gains in learners' comprehensibility compared to instruction targeting individual segments. It appears, then, that an instructional focus on L2 prosody and fluency may lead to a greater impact on comprehensibility than a focus on individual segments (see also Derwing *et al.*, 2014). In another study, Isaacs and Trofimovich (2012) examined how various segmental, prosodic and temporal characteristics of L2 speech (18 speech measures in total) interact to determine the comprehensibility of 40 native French speakers of English. Their findings showed that word stress (prosody) distinguished speakers of low, mid and high levels of comprehensibility, while speech rate (fluency) discriminated between low and intermediate levels, and vowel and consonant errors (segmental accuracy) distinguished intermediate from high levels. Similar findings were reported in follow-up studies featuring 60 ESL learners from various native language (L1) backgrounds (Crowther *et al.*, 2015b) and 120 Japanese learners of English (Saito *et al.*, 2016).

A growing number of studies have recently focused on vocabulary–comprehensibility links, targeting lexical profiles of advanced, intermediate and beginner level learners' spontaneous production. In these studies, L2 speech is often evaluated from written transcripts rather than from audio-recordings, to minimize pronunciation and fluency influences on speech assessments. For example, transcript-based ratings of lexical proficiency (ranging from 'high' to 'low') have been shown to be related to lexical sophistication (in terms of word frequency counts), abstractness (measured as lexical hierarchy), and lexical appropriateness (defined through collocation

accuracy) (Crossley *et al.*, 2011, 2015). In our own recent study (Saito *et al.*, in press), we also asked raters to judge comprehensibility in transcribed L2 speech samples. We found that lexical appropriateness (number of lexical errors), variation (lexical diversity), fluency (number of fillers produced), and abstractness (word imageablity) were crucial for distinguishing between beginner and intermediate comprehensibility levels. When it came to advanced comprehensibility, raters also seemed to attend to morphological appropriateness (morphology errors) and were sensitive to the use of semantically complex words with multiple senses.

Motivation for the current study

Apart from the linguistic characteristics of the speech itself, other variables can influence L2 comprehensibility. One source of such influences relates to various listener characteristics, which include the amount of listeners' exposure to and experience with L2 speech, the extent of their linguistic training, or the degree to which their own language background overlaps with that of the speaker. For example, listeners who are familiar with accented L2 speech or those who share the speakers' L1 tend to rate L2 speech differently, demonstrating more lenient attitudes towards accented speech, compared to listeners without relevant experience (e.g. Kennedy & Trofimovich, 2008; Winke *et al.*, 2013). Prior linguistic training also appears to matter for speech ratings. As shown by Isaacs and Thomson (2013), unlike novice raters who do not possess sufficient knowledge to articulate their rating decisions, experienced raters (ESL professionals) can explain their judgements by drawing on their linguistic knowledge and access to vocabulary or applied linguistics jargon with which to express themselves.

While it is clear that listeners' characteristics influence how they evaluate global aspects of L2 performance (e.g. in terms of overall accent or comprehensibility), the extent to which listeners with different experience profiles attend to similar linguistic characteristics of L2 speech to arrive at their rating decisions is unknown. This issue is important because understanding expert and novice judges' rating processes can inform the training of raters, particularly in the context of high-stakes language proficiency tests (e.g. IELTS), where all participating raters are expected to demonstrate consistent L2 speech assessments (Winke *et al.*, 2013). Recently, Saito and Shintani (2016) took an exploratory approach towards examining linguistic correlates of L2 comprehensibility, as perceived by listeners from different backgrounds (Singaporean and Canadian raters). The Singaporean raters, who had access to various native and nonnative models of English and also spoke a few L2s in a multilingual environment, tended to assign more lenient comprehensibility judgements compared to raters in Canada. Singaporean raters paid attention not only to pronunciation aspects of L2 speech but also to its lexical and grammatical content. In contrast, the comprehensibility

judgements of the Canadian raters, who used only North American English in a monolingual environment, were mainly determined by the pronunciation accuracy and fluency of L2 speech. However, the raters in Saito and Shintani's study evaluated relatively short samples (30 seconds), which may have been too short to capture various linguistic aspects of L2 speech (especially those specific to its lexis and grammar content) so that their relative contribution to L2 comprehensibility could be determined.

The current study extended previous research investigating rater influences on L2 speech assessment (Isaacs & Thomson, 2013; Saito & Shintani, 2016) by focusing on expert and novice raters' assessments of L2 comprehensibility. Two separate rater sessions with expert and novice raters were conducted to examine the role of pronunciation and lexis in L2 comprehensibility. In the first session, the raters evaluated audio samples so that their ratings could be related to extensive analyses of the same samples for several pronunciation variables (i.e. segmental and syllable structure errors, word stress, intonation and speech rate). In the second session, they evaluated transcribed speech, and their ratings were compared to extensive lexical analyses of the same speech samples (i.e. in terms of frequency, diversity, polysemy, hypernymy, text length, lemma and morphology). In line with previous L2 vocabulary research (e.g. Crossley *et al.*, 2011, 2015), the targeted speech samples were relatively long (about 3 minutes), which maximized the likelihood that they included a variety of pronunciation and lexical phenomena that could be linked to L2 comprehensibility.

Pronunciation Aspects of Comprehensibility

Rating materials

The speech samples consisted of 40 native French speakers' descriptions of an eight-frame cartoon narrative from our previous research (e.g. Isaacs & Trofimovich, 2012; Saito *et al.*, 2015). The speakers represented a range of L2 speaking ability, from complete beginners to simultaneous French-English bilinguals. The length of the recorded audio samples varied from 55 to 351 seconds (M = 146 seconds), which corresponded to 75–485 words produced (M = 209.2).

Audio-based comprehensibility analyses

Expert and novice raters

We recruited: (a) five expert raters who were graduate students in applied linguistics at an English-speaking university in Montreal, Canada (Mage = 28.0); and (b) five novice raters who were undergraduate or graduate students with non-linguistic majors at the same school (Mage = 22.6).

As residents in a bilingual (French-English) city, the raters were comparable in terms of their high familiarity with French-accented English (Kennedy & Trofimovich, 2008). However, in line with Isaacs and Thomson's (2013) definition of experienced and inexperienced raters, the two rater groups differed in their familiarity with L2 speech assessment. The expert raters had all taken graduate-level linguistics courses where they had received training in pronunciation, vocabulary and grammar analyses. These raters additionally reported on average 3.5 years (8 months–12 years) of prior English teaching experience, where they were tasked with evaluating their own students' L2 proficiency. In contrast, the novice raters had not completed any courses in linguistics and had no experience with teaching English.

Procedure

Following Derwing and Munro (2009: 478), comprehensibility was defined as 'the listener's perception of how easy or difficult it is to understand a given speech sample', and measured via scalar judgements. As described in Saito *et al.* (2015), the raters used a moving slider to provide a comprehensibility score on a scale between 0 = 'hard to understand' and 1000 = 'easy to understand'. If the slider was placed at the leftmost end of the continuum, labelled with a frowning face (indicating the negative endpoint), it was recorded as 0; if it was placed at the rightmost end of the continuum, labelled with a smiley face (indicating the positive endpoint), it was recorded as 1000. The raters first received brief instruction from a trained research assistant (via training scripts and onscreen labels, as shown in the Appendix). After familiarizing themselves with the rating procedure by rating three practice samples, they proceeded to the main dataset, with all 40 samples played randomly through a MATLAB interface. To ensure that raters' judgements reflected their intuitions, resembling real-life experiences with speech, the raters listened to each sample only once but were required to listen to each sample in its entirety. To reduce fatigue, the rating took place in two one-hour sessions.

Pronunciation analyses

As reported in Isaacs and Trofimovich (2012), the speech samples were analyzed for five pronunciation variables, with all analyses carried out and verified by two trained coders. The intraclass correlations were > 0.90. The five pronunciation variables were operationalized as follows:

(1) *Segmental error ratio*, defined as the total number of phonemic substitutions (e.g. 'put' spoken with /u/ in place of /ʊ/) divided by the total number of segments articulated.
(2) *Syllable structure error ratio*, defined as the total number of vowel and consonant epenthesis (insertion) and elision (deletion) errors (e.g. 'have' spoken without the initial /h/) divided by the total number of syllables articulated.

(3) *Word stress error ratio*, defined as the total number of instances of word stress errors (misplaced or missing primary stress) in polysyllabic words (e.g. 'SUIT-case' spoken as 'suit-CASE') over the total number of polysyllabic words produced.
(4) *Intonation error ratio*, defined as the number of correct pitch patterns at the end of phrases (syntactic boundaries) over the total number of instances where pitch patterns were expected (e.g. 'In a busy street [level tone], there is a businessman and a businesswoman [falling tone]').
(5) *Articulation rate*, defined as the total number of syllables produced excluding dysfluencies (e.g. filled pauses, repetitions, self-corrections, false starts) over the total speech sample duration.

Results

We first calculated Cronbach's alpha to check inter-rater agreement in raters' comprehensibility judgements. The expert raters showed higher consistency ($\alpha = 0.91$) than the novice raters ($\alpha = 0.81$). Since these indexes exceeded benchmark consistency values ($\alpha = 0.70$; Larson-Hall, 2010), mean comprehensibility ratings for each L2 speaker were computed by pooling the data across the five expert and five novice raters, respectively (see Table 8.1 for descriptive statistics).

Next, we compared the expert and novice raters' comprehensibility scores using a matched-samples *t*-test, which showed that the expert raters assigned significantly higher (and thus more lenient) comprehensibility scores compared to the novice raters ($t(39) = 3.05$, $p = 0.004$, $d = 0.21$). Finally, we examined how the expert and novice raters' comprehensibility scores were related to the five pronunciation variables in L2 speakers' speech, using correlation and regression analyses. As summarized in Table 8.2, correlation analyses showed that both expert and novice raters' comprehensibility scores were significantly associated with segmental, word stress and intonation errors, and nearly to the same degree.

We then performed two sets of multiple regression analyses to explore the degree to which the three pronunciation variables (segmental, word stress and intonation errors) predicted the expert and novice raters' comprehensibility scores. These analyses (summarized in Table 8.3) revealed that

Table 8.1 Descriptive statistics for expert and novice raters' comprehensibility scores on a 1000-point scale

Speaking dimension	*Mean*	*SD*	*Range*
Comprehensibility (expert)	713	196	267–998
Comprehensibility (novice)	667	233	214–1000

Note: 0 = 'hard to understand', 1000 = 'easy to understand'.

Table 8.2 Pearson correlations between the five pronunciation variables and expert and novice raters' comprehensibility ratings

Pronunciation variable	*Comprehensibility*	
	Expert raters	*Novice raters*
Segmental errors	–0.51*	–0.51*
Syllable structure errors	–0.36	–0.36
Word stress errors	–0.80*	–0.76*
Intonation errors	–0.51*	–0.51*
Articulation rate	0.32	0.38

Note: *$p < 0.01$ (Bonferroni adjusted).

Table 8.3 Results of multiple regression analyses using pronunciation variables as predictors of comprehensibility

Predicted variable	*Predictor variable*	*Adj. R^2*	*R^2 change*	*F*	*p*
Comprehensibility (expert)	Word stress	0.63	0.63	66.02	0.001
Comprehensibility (novice)	Word stress	0.56	0.56	50.78	0.001

Note: The variables entered into the regression included segmental, word stress and intonation errors; no evidence of strong collinearity was found (*VIF* < 1.259).

the number of word stress errors was the only significant predictor of the expert and novice raters' comprehensibility scores (accounting for a total of 62.5% and 56.1% of shared variance, respectively).

Lexical Aspects of Comprehensibility

Rating materials

To examine lexical contributions to expert and novice raters' comprehensibility judgements, the speaking materials used in the pronunciation analyses were transcribed and then rated by novice and expert raters for comprehensibility and analyzed for seven lexical variables.

Transcript-based comprehensibility analyses

Expert and novice raters

Following the same criteria used in the first analysis, we recruited five expert and five novice raters (*M*age = 29.3 years). None of these raters was involved in the investigation of the pronunciation aspects of comprehensibility. The expert raters (graduate students in applied linguistics) reported

having linguistic training and familiarity with pronunciation, vocabulary and grammar analyses, as well as a mean of 5.2 years of language teaching experience (2–10 years). The novice raters had not taken any linguistic courses nor taught language and thus had never experienced formal assessment of learner language.

Procedure

As with the previous analysis, the raters first received a brief explanation of comprehensibility (i.e. defined as effort in understanding what someone is trying to convey) from a trained research assistant (see Appendix to this chapter). Then the raters practised by evaluating three sample transcripts (not included in the main dataset), after which they proceeded to evaluate the 40 target transcripts. The transcripts were randomly presented on a computer screen through a MATLAB interface, and the raters used a free-moving slider to assess comprehensibility on a scale between 0 = 'hard to understand' and 1000 = 'easy to understand'. To ensure that the raters paid close attention to the transcripts, they were only allowed to make their judgements after spending at least five seconds reading each transcript.

Lexical analyses

Following Saito *et al.* (2015), the transcripts were analyzed for five lexical variables using the *Coh-Metrix* software (Graesser *et al.*, 2004) and for two additional variables (lexical appropriateness and morphological accuracy) through the coding of two trained coders. The intra-class correlations were beyond 0.90. The seven lexical variables were operationalized as follows:

(1) *Frequency* was calculated as the average frequency of vocabulary in the texts, using the word frequency scores included in the CELEX Lexical Database.
(2) *Diversity*, defined as 'the range and variety of vocabulary deployed in a text by either a speaker or a writer' (McCarthy & Jarvis, 2007: 459) was calculated using McCarthy's (2005) Measure of Textual Lexical Diversity (MTLD). MTLD derives diversity scores that are mathematically adjusted for varied text length (McCarthy & Jarvis, 2010).
(3) *Polysemy* was defined as the number of related senses in a single lexical entry. For example, 'man' has several meanings, such as 'an adult male person', 'humankind', 'husband', 'a male lover' and 'a subordinate'. Yet 'car' has fewer meanings, and these are primarily limited to either 'automobile' or 'a vehicle running on rails'.
(4) *Hypernymy* was defined as the hierarchical connections between general and specific lexical items, which facilitate the efficient processing and generalization of word knowledge. For example, words like 'transportation' and 'parents' are considered to be more general and less specific than words like 'car' and 'mother'.
(5) *Text length* was defined as the total number of words in each text.

(6) *Lexical appropriateness* was defined as the number of inaccurate and inappropriate words used, including L1 substitutions.
(7) *Morphological accuracy* was defined as the number of morphological errors including verb (i.e. tense, aspect, modality and subject-verb agreement), noun (i.e. plural usage related to countable and uncountable nouns), derivation (i.e. wrong derivational forms, such as 'surprised' instead of 'surprise'), and article (i.e. article usage in terms of finite, infinite and non-articles, and possessive determiners) errors.

Results

Analyses of rater consistency (Cronbach's alpha) revealed higher agreement for the expert ($\alpha = 0.93$) than for the novice raters ($\alpha = 0.86$). Again, because these indices exceeded the threshold of rating consistency typically assumed to be acceptable ($\alpha = 0.70$; Larson-Hall, 2010), the comprehensibility scores for each speaker were averaged across the expert and novice raters, respectively (see Table 8.4 for descriptive statistics). A comparison of the expert and novice raters' comprehensibility scores using a paired-samples *t*-test showed that the expert raters assigned significantly higher (more lenient) comprehensibility scores, compared to the novice raters ($t(39) = 3.104$, $p = 0.004$, $d = 0.23$).

We also performed correlation analyses to explore the relationship between the expert and novice raters' comprehensibility judgements and the seven lexical variables in L2 speakers' speech. As summarized in Table 8.5, comprehensibility scores were associated with the diversity, polysemy and lexical appropriateness variables for both groups of raters. However, a significant link between the morphological accuracy and comprehensibility variables was found only among the expert raters.

The four lexical variables significantly associated with comprehensibility were subsequently submitted to multiple regression analyses to examine the extent to which these variables predicted the expert and novice raters' comprehensibility ratings. Both the expert and novice raters' comprehensibility scores were equally predicted by the lexical appropriateness and diversity measures (Table 8.6). However, lexical appropriateness explained much of the variance in the expert raters' scores (71%), whereas diversity accounted for most of the variance in the novice raters' judgements (50%).

Table 8.4 Descriptive statistics for expert and novice raters' comprehensibility scores on a 1000-point scale

Speaking dimension	*Mean*	*SD*	*Range*
Comprehensibility (expert)	633	263	87–987
Comprehensibility (novice)	575	235	72–952

Note: 0 = 'hard to understand', 1000 = 'easy to understand'.

Table 8.5 Pearson correlations between the seven lexical variables and expert and novice raters' comprehensibility ratings

Lexical variable	*Comprehensibility*	
	Expert raters	*Novice raters*
Frequency	0.25	0.38
Diversity	0.55*	0.47*
Polysemy	0.57*	0.49*
Hypernymy	0.20	0.04
Text length	0.20	0.03
Lexical appropriateness	0.84*	0.71*
Morphological accuracy	0.52*	0.39

Note: *$p < 0.01$ (Bonferroni adjusted).

Table 8.6 Results of multiple regression analyses using lexical variables as predictors of comprehensibility

Predicted variable	*Predictor variable*	*Adj. R^2*	*R^2 change*	*F*	*p*
Comprehensibility (expert)	Appropriateness	0.71	0.71	63.47	0.001
	Diversity	0.77	0.06		
Comprehensibility (novice)	Diversity	0.50	0.50	50.78	0.001
	Appropriateness	0.64	0.14		

Note: The variables entered into the regression included diversity, polysemy, lexical appropriateness and morphological accuracy; no evidence of strong collinearity was found (*VIF* < 1.259).

Discussion

The current study was designed to examine whether and to what degree expert and novice raters (i.e. raters with linguistic and pedagogic backgrounds versus raters without professional experience in L2 classroom teaching) perceive the comprehensibility of L2 speech as a function of its pronunciation and lexical content. The global analyses showed that the expert raters assigned higher (more lenient) comprehensibility scores than the novice raters when evaluating both audio samples and transcripts of speech. These findings are in line with previous L2 speech research which shows that raters with L2 teaching experience and/or enhanced familiarity with particular L2 accents tend to be more lenient in their assessments of L2 speech relative to untrained teachers who have less exposure to accented speech (e.g. Isaacs & Thomson, 2013; Kennedy & Trofimovich, 2008; Winke *et al.*, 2013).

Additionally, the pronunciation and lexical analyses of L2 speech revealed that the expert and novice raters attended to overlapping yet somewhat distinct linguistic dimensions of L2 speech in rating comprehensibility. With respect to pronunciation variables, both expert and novice raters similarly relied on acoustic-phonetic information in L2 speech, in this case prioritizing the prosodic factor (word stress) over segmental accuracy (Crowther *et al.*, 2015a; Derwing & Munro, 2009; Derwing *et al.*, 1998; Field, 2005; Isaacs & Trofimovich, 2012; Kang *et al.*, 2010). With respect to lexical variables, the two sets of raters also seemed to attend to comparable domains of L2 vocabulary use, such as diversity (Koizumi & In'nami, 2012), polysemy (Crossley *et al.*, 2010), lemma appropriateness (Crossley *et al.*, 2015) and morphological accuracy (Yuan & Ellis, 2003). However, the relative weights of these lexical influences differed between the expert and novice raters. Unlike the novice raters, whose comprehensibility judgements were primarily linked to lexical diversity, the expert raters attended not only to how many different words L2 speakers used but also to whether they used them in a contextually appropriate manner.

In essence, these findings support Saito and Shintani's (2016) suggestion that more experienced raters' leniency towards L2 speech may be attributed to their sensitivity to, in particular, lexical content of L2 speech. More specifically, the expert raters seem to make a greater effort to understand what L2 speakers intend to convey, at least in terms of the lexical composition of speech, perhaps despite the fact that some of the spoken words are used contextually and conceptually inappropriately. In contrast, the less experienced raters appear to attend to surface-level L2 lexical characteristics such as lexical diversity, and focus less on the appropriateness of word use, which would make understanding of L2 speech more effortful. This difference in rater behaviour could be attributed to the expert raters' L2 teaching experience as language teachers, as well as to their expertise in applied linguistics (Isaacs & Thomson, 2013).

Implications for Second Language Assessment

The findings in this study offer several implications for rater training, particularly in high-stakes assessment contexts targeting the evaluation of L2 proficiency, where all raters should have an understanding of possible sources of rater bias to minimize individual differences among potentially heterogeneous raters (Xi & Mollaun, 2011). For example, raters with little linguistic or teaching experience could be informed that experienced raters judge L2 speech by attending not only to form (pronunciation and diversity), but also to meaning (appropriateness of word use). Based on previous research targeting listener recognition of L2 speech (Bradlow & Bent, 2008), it is possible that exposing raters with little linguistic or teaching experience to a

variable, diverse set of L2 speech (e.g. in terms of accents, speech rates or proficiency levels) can improve rater consistency in speech assessment, particularly with respect to L2 comprehensibility.

Since successful L2 communication can (and should be) treated as a consequence of joint action between the speaker and the listener (Jenkins, 2000; Levis, 2005), it is noteworthy that much research to date has largely focused on the L2 speaker, highlighting problematic areas in need of improvement in terms of their pronunciation. Relatively few studies have examined how listeners should accommodate their listening strategies to better understand accented L2 speech (see Derwing *et al.*, 2002; Jenkins, 2000; Kang *et al.*, 2014). Assuming that listeners' assessments of L2 accent are largely based on pronunciation aspects of L2 speech, while their evaluations of L2 comprehensibility draw on a variety of linguistic dimensions (Crowther *et al.*, 2015b; Isaacs & Trofimovich, 2012; Saito *et al.*, 2016), raters might need to be told that L2 comprehensibility ratings capture listeners' ability to extract word- and discourse-level meaning from L2 speech. To avoid being distracted by nonnative pronunciation patterns, raters might also need to be made aware of perceptually salient characteristics of L2 speech which do not necessarily hinder understanding. As such, raters can focus on evaluating the comprehensibility of their speech without penalizing L2 speakers for their nonnative-like use of phonological features with little communicative value (Derwing & Munro, 2009), such as segments with low functional load (Munro & Derwing, 2006), *schwa* insertion in complex syllables (Lin, 2003), and monotonous (but not necessarily erroneous) prosody (Jenkins, 2000).

Limitations

Due to the exploratory nature of this study, several limitations need to be acknowledged. One obvious limitation is the small sample size of L2 speakers (40) limited to a single linguistic background (French), and native-speaking raters (10 for audio- and transcribed-based comprehensibility analyses, respectively). Secondly, this study focused on only one rater characteristic, namely raters' experience with L2 assessment through graduate-level linguistic training and/or language teaching. Thus, it would be important to examine how other rater background variables, such as the amount of familiarity with L2 speech (Kennedy & Trofimovich, 2008) and L2 learning background (Winke *et al.*, 2013), can influence raters' sensitivity to linguistic information during L2 speech assessment. Thirdly, the current findings were based on raters' judgements of L2 speech elicited via a single task (picture description). Because the same L2 users' speaking performance tends to vary (e.g. in terms of linguistic complexity, accuracy and fluency) across tasks (Robinson, 2011), future research needs to examine how rater experience influences the assessment of L2 speech elicited under various task conditions, including the

availability of planning time (Yuan & Ellis, 2003), task repetition (Ahmadian & Tavakoli, 2011), story complexity (Tavakoli & Foster, 2010), and the presence or absence of an interlocutor (Crowther *et al.*, 2015a). Finally, the study only relies on quantitative data and, thus, is not able to probe rater perceptions and triangulate these with correlations between listener-coded measures and the scores they assign, as in the Isaacs and Thomson's (2013) study. Also, it is unclear whether the measures that were identified for the study are actually those that raters attend to during normal operational ratings in research contexts.

Conclusion

The current study investigated the role of rater experience in listener-based judgements of L2 comprehensibility, focusing on two groups of native-speaking raters with and without classroom teaching experience. Results showed that expert raters (graduate students in applied linguistics and teaching professionals) provided more lenient comprehensibility ratings than novice raters. Secondly, the study demonstrated that raters with and without professional experience in L2 teaching and (by implication) experience in assessment were both similar and different in the extent to which they relied on various linguistic dimensions of L2 pronunciation in relation to comprehensibility. While both expert and novice raters processed pronunciation information in a comparable fashion (by drawing particularly on prosody), they revealed different patterns of behaviour with regard to lexical dimensions of speech. For novice raters, comprehensibility was linked to the number of different words used by L2 speakers; for expert raters, comprehensibility was largely associated with the appropriateness of word use. Building on previous comprehensibility research (e.g. Derwing & Munro, 2009; Isaacs & Trofimovich, 2012) as well as rater-focused studies (e.g. Saito & Shintani, 2016; Winke *et al.*, 2013), the current findings highlight the importance of studying the complex relationship between rater background, linguistic composition of speech, and L2 comprehensibility, with the goal of improving both the success of L2 communication and a better understanding of the linguistic constructs being measured in order to enhance the validity of the assessment.

References

Ahmadian, M.J. and Tavakoli, M. (2011) The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research* 15, 35–59.

Bradlow, A.R. and Bent, T. (2008) Perceptual adaptation to non-native speech. *Cognition* 106, 707–729.

Crossley, S.A., Salisbury, T. and Mcnamara, D.S. (2010) The development of polysemy and frequency use in English second language speakers. *Language Learning* 60, 573–605.

Crossley, S.A., Salsbury, T., Mcnamara, D.S. and Jarvis, S. (2011) What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly* 45, 182–193.
Crossley, S.A., Salsbury, T. and Mcnamara, D.S. (2015) Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics* 36, 570–590.
Crowther, D., Trofimovich, P., Isaacs, T. and Saito, K. (2015a) Does speaking task affect second language comprehensibility? *Modern Language Journal* 99, 80–95.
Crowther, D., Trofimovich, P., Saito, K. and Isaacs, T. (2015b) Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly* 49, 814–837.
Derwing, T. and Munro, M. (1997) Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition* 12, 1–16.
Derwing, T. and Munro, M. (2005) Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly* 39, 379–397.
Derwing, T.M. and Munro, M.J. (2009) Putting accent in its place: Rethinking obstacles to communication. *Language Teaching* 42, 476–490.
Derwing, T., Munro, M. and Wiebe, G. (1998) Evidence in favor of a broad framework for pronunciation instruction. *Language Learning* 48, 393–410.
Derwing, T.M., Munro, M.J. and Rossiter, M.J. (2002) Teaching native speakers to listen to foreign-accented speech. *Journal of Multilingualism and Multicultural Development* 23, 245–259.
Derwing, T.M., Rossiter, M.J., Munro, M.J. and Thomson, R.I. (2004) L2 fluency: Judgments on different tasks. *Language Learning* 54, 655–679.
Derwing, T.M., Munro, M.J., Foote, J.A., Waugh, E. and Fleming, J. (2014) Opening the window on comprehensible pronunciation after 19 years: A workplace training study. *Language Learning* 64, 526–448.
Field, J. (2005) Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly* 39 (3), 399–423.
Graesser, A.C., Mcnamara, D.S., Louwerse, M.M. and Cai, Z. (2004) Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers* 36, 193–202.
Hahn, L.D. (2004) Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly* 38, 201–223.
Isaacs, T. and Thomson, R.I. (2013) Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly* 10, 135–159.
Isaacs, T. and Trofimovich, P. (2012) Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition* 34, 475–505.
Jenkins, J. (2000) *The Phonology of English as an International Language*. Oxford: Oxford University Press.
Kang, O., Rubin, D. and Pickering, L. (2010) Suprasegmental measures of accentedness and judgments of English language learner proficiency in oral English. *Modern Language Journal* 94, 554–566.
Kang, O., Rubin, D. and Lindemann, S. (2014) Mitigating U.S. undergraduates' attitudes toward international teaching assistants. *TESOL Quarterly* 49, 681–706. doi:10.1002/tesq.192.
Kennedy, S. and Trofimovich, P. (2008) Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review* 64, 459–489.
Koizumi, R. and In'nami, Y. (2012) Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System* 40, 554–564.

Larson-Hall, J. (2010) *A Guide to Doing Statistics in Second Language Research Using SPSS*. New York: Routledge.

Levis, J. (2005) Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly* 39, 367–377.

Lin, Y. (2003) Interphonology variability: Sociolinguistic factors affecting L2 simplification strategies. *Applied Linguistics* 24, 439–464.

McCarthy, P.M. (2005) An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). Unpublished PhD dissertation, University of Memphis.

McCarthy, P.M. and Jarvis, S. (2007) vocd: a theoretical and empirical evaluation. *Language Testing* 24, 459–488.

McCarthy, P.M. and Jarvis, S. (2010) MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42, 381–392.

Munro, M. and Derwing, T. (1995) Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning* 45, 73–97.

Munro, M. and Derwing, T. (2006) The functional load principle in ESL pronunciation instruction: An exploratory study. *System* 34, 520–531.

Robinson, P. (2011) *Second Language Task Complexity: Researching the Cognition Hypothesis of Language Learning and Performance*. Amsterdam: John Benjamins.

Saito, K. and Shintani, N. (2016) Do native speakers of North American and Singapore English differentially perceive second language comprehensibility? *TESOL Quarterly* 50, 421–446.

Saito, K., Trofimovich, P. and Isaacs, T. (2015) Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, 29 September. doi:10.1093/applin/amv047.

Saito, K., Trofimovich, P. and Isaacs, T. (2016) Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics* 37, 217–240.

Saito, K., Webb, S., Trofimovich, P. and Isaacs, T. (in press) Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness and sense relations. *Studies in Second Language Acquisition* 38.

Tavakoli, P. and Foster, P. (2010) Task design and second language performance: The effect of narrative type on learner output. *Language Learning* 58, 439–73.

Trofimovich, P. and Isaacs, T. (2012) Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition* 15, 905–916.

Winke, P., Gass, S. and Myford, C. (2013) Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing* 30, 231–252.

Xi, X. and Mollaun, P. (2011) Using raters from India to score a large-scale speaking test. *Language Learning* 61, 1222–1255.

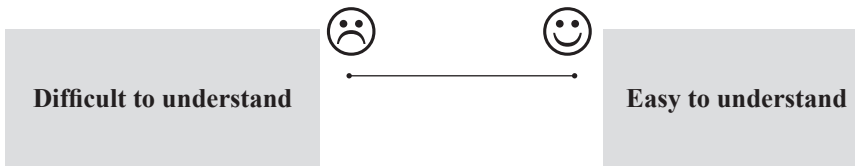
Yuan, F. and Ellis, R. (2003) The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics* 24, 1–27.

Appendix: Training Materials and Onscreen Labels for Comprehensibility Judgement

Training script

Comprehensibility refers to how much effort it takes to understand what someone is trying to convey. If you can understand (what the picture story is all about) with ease, then the speaker is highly comprehensible. However, if you struggle and must read very carefully, or in fact cannot understand what is being said at all, then a speaker has low comprehensibility.

Onscreen labels



9 Assessing Second Language Pronunciation: Distinguishing Features of Rhythm in Learner Speech at Different Proficiency Levels

Evelina Galaczi, Brechtje Post, Aike Li, Fiona Barker and Elaine Schmidt

Introduction

Ongoing globalization in the 21st century, as well as the primary role of English as a lingua franca in educational and professional contexts, has brought to the forefront the fundamental role of second/foreign language (L2) pronunciation, due to the growing need for interlocutors from different first language (L1) backgrounds to engage in meaningful and intelligible communication in English. While pronunciation is clearly an important skill, and insufficient pronunciation ability is detrimental to the intelligibility of speech, it has a surprisingly minor role in theoretical L2 models, in assessment scales of speaking, and in practical assessment and teaching contexts involving examiners, teachers and learners. In terms of theoretical models, pronunciation ability is only minimally dealt with in the widely accepted framework of communicative language ability (Bachman, 1990), where pronunciation (phonology) is neglected and categorized together with graphology, despite the fundamental difference between these two constructs (Isaacs, 2014). The backstage role of pronunciation in theoretical models is also reflected in some well-known assessment scales, where pronunciation tends to be captured inconsistently or to be entirely absent. For example, the ACTFL Proficiency Guidelines for Speaking (ACTFL, 2012) do not include descriptors for pronunciation at every level. Similarly, the Common European Framework of

Reference for Languages (CEFR) omits pronunciation as a criterion from the Spoken language use scale (Council of Europe, 2001: 28–29), attributing this decision to the difficulty in distinguishing across pronunciation ability levels in the same way as other language skills and the difficulty in interpreting descriptors of pronunciation consistently across languages (North & Hughes, 2003: 6). Pronunciation is, instead, captured in the phonological control scale (Council of Europe, 2001: 117), which includes only brief descriptors. It does not consistently describe the construct of interest, and does not distinguish between CEFR Levels C1 and C2 (see Harding, this volume).

The underdeveloped nature of the pronunciation construct is also seen in the challenge presented by pronunciation for teachers and examiners: studies have indicated that examiners find the assessment of pronunciation to be more challenging than that of other skills, and that they tend to be more confident when making global judgements of intelligibility than when making judgements about micro-level segmental and prosodic features (e.g. Brown & Taylor, 2006; Isaacs *et al.*, 2015; Yates *et al.*, 2011). Teachers have also been found to lack training and confidence in their pronunciation expertise (Levis, 2006).

The under-researched nature of the pronunciation construct is reflected in L2 acquisition research as well, where relatively little is known about how segmental and prosodic features develop over time. What is known is that cross-linguistic differences in segmental and prosodic proficiency are apparent due to language transfer (e.g. difficulties with the /l/ and /r/ contrast for Japanese learners), especially at lower levels of proficiency, reflecting properties of the L1 (Major, 2008). While the L1 influence is important, research has also indicated that acquisition does not necessarily proceed in a uniform fashion, and that some features are subject to L1 transfer while others, such as accentual lengthening, show common developmental paths across languages (e.g. Li & Post, 2014). The picture is further complicated by the fact that proficiency emerges with the acquisition of phonology, morphosyntax and information structure and the mapping between them (Post *et al.*, 2010) but phonological and prosodic acquisition can be out of step with acquisition in other areas of language competence at higher levels.

To sum up, there are difficulties with the theoretical conceptualization and practical operationalization of L2 pronunciation. One useful line of enquiry to pursue in making pronunciation a well-understood and integral part of learning, teaching and assessment contexts is to better understand the pronunciation features of learners with different L1s at different proficiency levels. The present study aims to contribute to this need through providing an in-depth micro-analytic investigation of prosodic features observed in learner speech. The prosodic features of interest relate to the rhythm of speech, chosen for investigation here largely due to the under-researched nature of rhythm in L2 speech, which is at odds with its important role in comprehensibility (Isaacs & Trofimovich, 2012). Through its dual

interdisciplinary phonetics and assessment perspective, the study aims to establish a profile of the rhythmic properties of learner speech at different proficiency levels, which can in turn contribute towards a more comprehensive definition and operationalization of the construct of L2 pronunciation. A broader aim is to raise awareness about micro-level features of rhythm and prosody which play a role in learner speech, and which teachers and assessors are likely to benefit from.

Role of Rhythm in English Speech

Rhythm has traditionally been seen as a key distinguishing feature between languages, with stress-timed and syllable-timed languages regarded as distinct on the basis of differences in rhythmic properties (Abercrombie, 1967; Pike, 1945). In stress-timed languages the durations between each stressed syllable tend to be approximately equal, whereas in syllable-timed languages the durations of syllables tend to be approximately equal. In stress-timed languages, stressed syllables are significantly longer than unstressed syllables, and unstressed syllables which occur between consecutive stressed syllables are compressed (e.g. through vowel reduction) to fit into the time interval. For example, in the phrases 'i LIKE to TRAvel' and 'i LIKE very much to TRAvel', the unstressed words/syllables (in lower case letters) are reduced and shorter than their stressed counterparts and, importantly, the intervals taken by the unstressed syllables are of approximately equal duration. In syllable-timed languages, in contrast, syllables have more equal duration and prominence, with little or no vowel reduction. Table 9.1 provides a summary of key properties that have been associated with the stress-timed versus syllable-timed rhythm classification.

Although the rhythmic distinctions between stress-timed and syllable-timed languages have been empirically supported (e.g. Ramus *et al.*, 2003), many instrumental studies have failed to find constant and systematic evidence to support the dichotomous approach to categorizing languages (e.g. Bolinger, 1965; Roach, 1982). In addition, it is clear that rhythm manifests itself along a number of phonetic dimensions, including duration, pitch and loudness. As a result, rather than categorizing languages in terms of two (or three) distinct rhythm classes, crosslinguistic differences in rhythm are now accounted for in terms of a continuous model of rhythm in which a combination of language-specific properties (besides more general factors such as speaking rate) result in different rhythmic patterning gradients along a continuum. Depending on these properties, an individual language will fall at a particular point along the continuum (e.g. Dauer, 1983; Prieto *et al.*, 2012). Bearing this in mind, we adopt the terms 'stress-timed' and 'syllable-timed' in this chapter for ease of reference, and only operationalize speech rhythm in terms of duration in the first instance.

Table 9.1 Stress-timed and syllable-timed languages: Rhythmic properties

Rhythmic property	*Stress-timed languages (e.g. Dutch, English, German)*	*Syllable-timed languages (e.g. French, Mandarin, Spanish)*
Vowel reduction Vowels in unstressed position tend to be shorter and to converge towards the central/ neutral vowel schwa /ə/.	Evident reduction of unstressed vowels.	Vowel reduction is not evident.
Syllable structure complexity Number of consonants allowed in a syllable.	Complex consonant clusters, therefore high amount of consonants in speech, e.g. C, CC and CCC.	Open syllables (CV) are far more common than complex syllable structures, therefore a lower amount of consonants in speech.
Durational marking of accentuation The lengthening of accented syllables compared to unaccented syllables.	Large durational difference between accented syllables and unaccented ones.	Little durational difference between accented and unaccented syllables.
Final lengthening The lengthening of phrase-final and utterance-final syllables compared to non-final syllables.	Final syllables are lengthened compared to non-final syllables.	Little durational distinction between final and non-final syllables.

Note: C = consonant; V = vowel.

Apart from such language-specific properties, rhythm is affected by non-linguistic factors such as speech rate (e.g. Prieto *et al.*, 2012). Figure 9.1 provides an illustration of how the consonants and vowels that are produced in the speech stream form intervals of different types with variable durations, which in turn create the rhythm of English speech. The example chosen is the fragment '… if you have a team leader, strictly speaking …' (see Figure 9.2 for the speech pressure wave and spectrogram). As Figure 9.1 shows, syllable boundaries do not necessarily coincide with the edges of words (e.g. in 'if you' the /f/ forms a syllable with the following word /ju:/), nor with the intervals that contain the sequences of vocalic or consonantal material (e.g. in 'strictly', the second consonantal interval straddles a syllable boundary). The figure also illustrates that the intervals are delimited by major phrase boundaries (marked by a silent pause # in this example). The effect of syllable structure can be seen when we compare intervals that contain complex

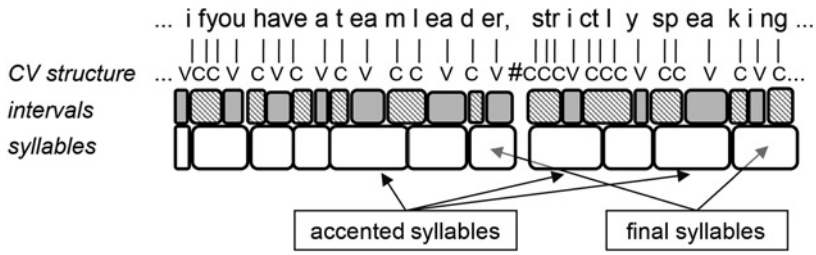


Figure 9.1 Stress timing: An example of syllable durations and vocalic and consonantal intervals

Notes: (i) C = consonant; V = vowel; grey boxes = vocalic intervals, striped boxes = consonantal intervals; # = silent pause. (ii) Vocalic interval = section of speech between vowel onset and offset; consonantal interval = section between vowel offset and onset. (iii) Durations are shown approximately to scale.

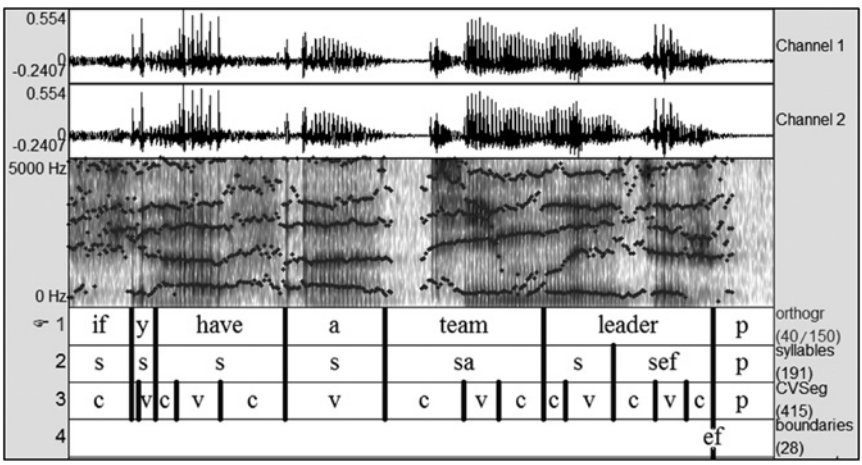


Figure 9.2 Example of the segmental and prosodic labelling

consonant clusters (e.g. /str/ in 'strictly') with those that contain singleton consonants (e.g. /t/ in 'team'). As a result, the duration of the consonantal intervals is quite variable, and will be much more variable than in a language that only allows syllables with singleton consonants. Also, accented syllables with full vowels tend to be considerably longer than unaccented syllables which often contain reduced vowels (e.g. /ɪ/ in 'strict', 'if' and '-ly'). The effect of boundary lengthening can be seen in the longer duration of syllables that precede a major phrase boundary (compare the duration of the schwa vowel in 'a' and '-der'). Although the durational marking of accented syllables and phrase boundaries is very common across the languages of the world, the amount of lengthening varies considerably, and Figure 9.1 exemplifies that that effect is particularly strong in English.

Rhythm Metrics

A number of metrics have been developed to quantify rhythm in (learner) speech (for a review, see White & Mattys, 2007).

%V (Proportion of vocalic material in speech)

Ramus *et al.* (1999) proposed several utterance-level measures of rhythm by dividing speech into vocalic and consonantal parts and computing the proportion of the vocalic interval duration in speech, expressed as a percentage from the total utterance duration: %V (i.e. comparing the amount of material in the grey against the striped boxes in Figure 9.1). In a stress-timed language such as English, %V would typically be lower due to vowel reduction and to the presence of consonant clusters, compared to a syllable-timed language such as Spanish.

In English, L2 learners may incorrectly insert vowels to break up consonant clusters, and as such would increase the proportion of vocalic material in their speech. For example, Japanese learners of English who are struggling with this pronunciation feature may pronounce the word 'Christmas', which has several consonant clusters, as /kurisumasu/. Learners may, in addition, have trouble with vowel reduction in unstressed positions, and make vowel intervals that contain unaccented syllables too long.

Varco-V and Varco-C (variability in vocalic or consonantal duration)

To measure the variability in vocalic or consonantal interval duration (i.e. quantifying the amount of variability within either the grey or the striped intervals in Figure 9.1), Dellwo (2006) developed Varco-C, which calculates the standard deviation of consonantal interval duration (normalized for speech rate), as well as its vocalic counterpart Varco-V. Varco measures the variability globally as an average across the whole speech of a speaker, in contrast to other measures (discussed below) which have a more local focus. Typically, this variability is expected to be larger in stress-timed languages, such as English, due to the higher variety in syllable structures associated with a stress-timed language (e.g. some syllables have complex consonant clusters and some simple structures), and greater accentual and final lengthening. In syllable-timed languages, in contrast, a large proportion of syllables have a simple CV structure and successive syllables are more similar in length, leading to lower variability values (Low, 2006).

In English, L2 speech may show lower variability in vocalic and consonantal durations, partly due to L1 transfer from a syllable-timed language. This will, in part, be due to processing and articulation difficulties associated with differences between the syllable structures in the L1 and L2, but also

with differences in accentual and boundary marking. In addition to possible transfer effects, accentuation has been shown to be correlated with proficiency level (Kang, 2013; Kang *et al.*, 2010) and with comprehensibility and accent ratings (Isaacs & Trofimovich, 2012) independent of the L1. Low-proficiency learners have been reported to overuse accentuation, since they accent lexical items regardless of their function or information load (Kang, 2010). This is likely to have knock-on effects on speech rhythm.

PVI (Pairwise Variability Index)

Other measures capture rhythm more locally by focusing on the degree of durational difference between neighbouring intervals. One such measure is the PVI (Pairwise Variability Index) metric which calculates the mean of the durational differences between successive temporal intervals in an intonation phrase (a stretch of speech with its own intonation contour). The raw PVI (rPVI-C for consonantal intervals and rPVI-V for vocalic intervals) contrasts with its normalized counterpart which controls for the effect of speech rate (nPVI-C and nPVI-V) (Grabe & Low, 2002; Low *et al.*, 2000). A stress-timed language would be expected to have high PVI values and a syllable-timed language low PVI values, as confirmed by Low *et al.* (2000) and Low (1998) who compared British English (stress-timed) and Singapore English (syllable-timed) and found that the British English speakers exhibited a significantly higher variability in duration between successive vowels than the Singapore English speakers. In English L2 speech, learners who are producing more syllable-timed speech would be expected to show lower variability in durations between neighbouring intervals.

Prosody, Rhythm and Second Language English Learners

Prosody, which comprises rhythm alongside intonation, tone, accentuation and boundary marking, has been empirically shown to play a fundamental role in comprehensibility (e.g. Anderson-Hsieh *et al.*, 1992; Kang *et al.*, 2010; Munro & Derwing, 1995; Pickering, 2001). In other words, a speaker with otherwise good articulation may be difficult to understand because of weak prosody. Stress – the most widely studied prosodic feature in L2 speech – has consistently been found to relate to L2 comprehensibility, both at word- and sentence-stress level (Field, 2005; Hahn, 2004; Kang *et al.*, 2010). Discussions in the area of English as a lingua franca have also indicated that communication breakdowns could be due to misplaced sentence stress (Jenkins, 2002).

Rhythm has been investigated less extensively than stress as a prosodic feature of L2 speech. Low (2006) suggests that the rhythm of a stress-timed

language would be more difficult to acquire than the rhythm of a syllable-timed language, due to the need to reduce vowels and compress syllables in stress-timed languages. This hypothesis finds support in various empirical studies: Gutiérrez-Díaz (2001) found that advanced Spanish learners of English produced English with a stressed/unstressed syllable durational ratio, which was mid-way between the ratios for Spanish and English, suggesting that stress-timing posed a problem for those learners. Similarly, Trofimovich and Baker (2006) reported a significant difference between inexperienced and moderately experienced learners and native speakers of English in terms of their stressed/unstressed syllable ratio, again showing that learners had difficulty producing stress timing. Vowel reduction – a fundamental component of rhythm in English – has also been related to measures and perceptions of accentedness and comprehensibility in learners (Trofimovich & Isaacs, 2012).

Other studies have included rhythm as part of investigations focusing on differences in learner speech across proficiency levels. In this respect, Iwashita *et al.* (2008), in an important first attempt to provide an in-depth empirical analysis of learner speech, focused on distinguishing levels of proficiency through a range of linguistic features in the context of the TOEFL iBT speaking test. The researchers utilized a range of pronunciation measures (alongside measures of grammatical accuracy and complexity, vocabulary and fluency) and focused on the pronunciation of word and sub-word level features, on intonation and on rhythm using auditory coding of features (e.g. targetlike/non-targetlike) and not instrumental measurements. A significant difference between levels was found only with the production of targetlike syllables, but the authors reported high correlations between rhythm and proficiency level, with appropriate rhythm associated with higher proficiency learners.

Kang (2013) analyzed a range of linguistic features at CEFR Levels B1 to C2 in a set of Cambridge English speaking test performances. The speech analysis program *Praat* (Boersma & Weenink, 2010) was used alongside listener coding of measures to analyze objective pronunciation measures, such as proportion of words with prominent stress, number of prominent syllables per run, overall pitch range and a range of tones. The findings indicated that there are objectively measured differences between high- and low-proficiency learners, but not necessarily between adjacent levels. The author further showed that, as proficiency increased, the proportion of stressed words within a sentence decreased, thus supporting prior research indicating that low-proficiency learners stress items regardless of their function or importance (Kang, 2013). In a similar line of research, Isaacs and Trofimovich (2012) set out to produce an empirically based rating scale for pronunciation and identified a subset of features which best distinguish between three levels of L2 comprehensibility, using both auditory and instrumental measures. The authors noted a strong relationship between word stress and

vowel reduction and raters' judgements in the sample, which, it needs to be noted, was limited to one L1 group (French).

The development of the rhythm metrics outlined in the previous section has given rise to investigations of L2 speech rhythm in a quantitative and systematic way. However, only a few studies have investigated rhythmic differences between L2 learners of different proficiency levels, or with different L1 backgrounds (e.g. Guilbault, 2002; Gut, 2009; see Li & Post, 2014, for review). These studies have provided inconsistent evidence for rhythmic differences in L2 speech depending on level of proficiency or L1.

Study Aim and Research Questions

As noted in the literature review, few studies have systematically examined how rhythm is displayed by L2 learners at different proficiency levels and from controlled L1 backgrounds, and even fewer studies to date have employed the rhythm metrics used here in the context of learner speech at different proficiency levels. The aim of this study, therefore, is to offer a more comprehensive empirically based investigation of rhythm in L2 speech than that given in previous studies in order to establish to what extent rhythmic measures can discriminate between proficiency levels. We do so through a small-scale quantitative investigation of objectively measurable micro-level prosodic rhythmic features in the speech of learners at different proficiency levels (reported as CEFR levels) and through controlling for learner L1 background. The study aims to provide more granularity in the analysis by moving from a judgement that a specific feature is not targetlike (e.g. Iwashita *et al.*, 2008), to an investigation of what makes it not targetlike at different CEFR levels. The main motivation is the need to better understand L2 pronunciation as a construct and to provide practical findings that can inform learning, teaching and assessment. A broader aim of the study is to contribute a cross-disciplinary perspective to L2 pronunciation through collaboration between L2 phoneticians and language assessment specialists. The following four research questions guide the study; the first two deal with pronunciation across proficiency levels while the last two deal with pronunciation across L1s.

(1) How reliably can levels of L2 pronunciation ability be discriminated across CEFR Levels A1–C2 using a set of rhythmic measures?
(2) Which rhythmic measures have the highest discriminative properties for particular proficiency levels?
(3) How far do rhythmic measures display different patterns for learners of different L1 backgrounds?
(4) Which rhythmic measures display the largest differences for particular L1s?

Methodology

Speech samples

Speech samples drawn from speaking test performances of 20 English learners from three typologically different L1 backgrounds and six CEFR levels were used in this study (see Table 9.2).

Approximately 60 seconds of speech were used per learner, taken from a Question-and-Answer task, where the learners responded to a series of questions and produced extemporaneous speech. The speech samples at CEFR Levels A2–C2 were taken from Cambridge English face-to-face speaking tests, whereas at Level A1 performances were extracted from a computer-delivered speaking test. In both test formats the same task type was used to minimize any differences due to a method effect.

The participants represented 'average' learners at each CEFR level based on their pronunciation score; that is, they were not borderline within their CEFR level. Borderline test takers with marks at the top or bottom of the scale would have been likely to show pronunciation features typical of the adjacent proficiency levels and were not deemed suitable for analysis. Rater effects were minimized by using Fair Average marks generated by Facets (Linacre, 1989) which were based on multiple marks from a group of accredited experienced examiners (marks were provided by Cambridge English).

Two studies served as the basis of the empirical investigation reported here. In Study 1, speech was used from 12 L1 Spanish speakers spanning Levels A1 to C2 (two learners at each level) to examine the variability in learner performances across proficiency levels, while controlling for L1 background; in Study 2, speech was used from L2 speakers from Korean, Spanish and German L1 (two each at Levels B1 and B2) to compare L1 effects. The three languages were chosen to be typologically different: German is stress-timed (Kohler, 1982); Spanish is syllable-timed (Pike, 1945); whereas Korean is generally considered to be neither stress-timed nor syllable-timed (Seong, 1995). CEFR Levels B1 and B2 formed the basis of Study 2.

Table 9.2 Dataset

Level	*L1*		
	German	*Spanish*	*Korean*
A1	–	2	–
A2	–	2	–
B1	2	2	2
B2	2	2	2
C1	–	2	–
C2	–	2	–

Speech measures

The 20 speech samples were analyzed for a number of measures which quantify different aspects of speech rhythm, and which were chosen based on previous research linking these measures to proficiency level or L1 differences. Table 9.3 provides a definition of these measures and their operationalization. Taken together, an analysis of learner speech across a range of proficiency levels using these measures should reveal how successful learners are at each level in producing speech that has stress-timing features.

Analysis

Syllable durations and durations of consonantal and vocalic intervals were extracted using *Praat* (Boersma & Weenink, 2010). An example of the segmental and prosodic labelling in *Praat* is given in Figure 9.2.

The first tier of the figure (fourth row down) contains orthographic transcription. The second tier (fifth row down) was used to calculate accentual and final lengthening. Each syllable is marked as unaccented and non-final (labelled as 's'), accented ('sa'), final ('sef'), accented and final ('safe') or hesitated ('sx'). Hesitated syllables were excluded from the analysis. The key segmentation criterion for syllabification of the speech produced by the study participants was Gussenhoven and Jacob's (2005) version of the Maximum Onset Principle (Pulgram, 1970), which would syllabify the word 'leader' as 'lea.der' instead of, for instance, 'lead.er'. Establishing this criterion was important to ensure consistency in the measurement of syllable durations across the different speech samples. The second tier was also used to calculate speech rate by counting the number of syllables that were realized in each 60-second speech sample. The third and fourth tier together provided the information required for the calculation of the rhythm metrics. In the third tier, each vowel and consonant was segmented primarily by visual inspection of the speech waveforms and wideband spectrograms with reference to standard criteria (e.g. Peterson & Lehiste, 1960; White & Mattys, 2007). The fourth tier contains the phrasing information, that is, the beginning and end of an intonation phrase. Within each intonation phrase, consecutive consonant/vowel intervals were merged into vocalic and consonantal intervals.

Inter-coder agreement was measured for a subset (15%) of the data (one speaker per language). The samples used for inter-coder estimation were segmented and labelled by three annotators. The inter-coder agreement was 97% (calculated as number of codes of agreement/total number of codes).

Once the relevant measures were derived, statistical analyses were performed using IBM *SPSS 20*. The data were analyzed using multivariate analysis of variance (MANOVA) with factors Language background (Spanish, German, Korean) and Proficiency level (A1, A2, B1, B2, C1, C2), and fixed factors Rhythm metrics (nPVI-V, Varco-V, Varco-C, %V, rPVI-C, nPVI-C).

Table 9.3 Speech measures used in this study

Measure		*Operationalization*
Speaking rate		Number of syllables in 60 sec sample.
Accentuation		Durational differences between accented and unaccented syllables. For example, in the sentence 'There are some CHILdren in this PHOto', the duration of the accented syllables (given in capital letters) should be relatively longer than the duration of the unaccented syllables. The appropriateness of accent placement was not considered; only the durational difference between accented and unaccented syllables was.
Boundary marking		Durational difference between phrase-final and non-phrase final syllables (in an intonation phrase). For example, in the sentence 'I LIKE playing drums', the duration of the final syllable (drums) in the intonation phrase should be longer than if it was followed by another word in the phrase, as in 'I LIKE playing drums at home.'
Rhythm metrics	%V	Proportion of vocalic material in speech. For example, the word 'stop' would have a lower proportion of vocalic material than the incorrectly pronounced /sətop/ with a schwa sound inserted to break up the consonant cluster. Function words, which typically have reduced vowels, may also add to the proportion of vocalic material in L2 speech if the vowels are not reduced.
	Varco-V and Varco-C	Variability in vocalic/consonantal interval duration (standard deviation divided by mean). In contrast to PVIs (below), this is averaged over all utterances. For example, the cluster /str/ in 'strength' as opposed to /t/ in 'teams'.
	rPVI-C	Durational differences between adjacent consonantal intervals in an intonation phrase (raw). In contrast to Varco, all variation here is measured locally, by comparing each interval with its successive interval. For example, the cluster /str/ in 'strength' as opposed to /t/ in 'teams'.
	nPVI-V	Durational differences between adjacent vocalic intervals in an intonation phrase (normalized for speaking rate). For example, in 'YESterday' the vocalic interval in 'yes' is much longer than in 'ter' (because of vowel reduction), followed by a longer vocalic interval in 'day'.
	nPVI-C	Durational differences between adjacent consonantal intervals in an intonation phrase (normalized for speaking rate). For example, the cluster /str/ in 'strength' as opposed to /t/ in 'teams'.

Results

Study 1: Differences across proficiency levels

We will first present the results of the 12 Spanish learners across six different proficiency levels, starting with the rhythm metrics. Of the six rhythm metrics of interest here, three (seen in Figure 9.3) were found to be discriminative for CEFR level: %V (proportion of vocalic material in speech), Varco-C (variability in consonantal interval duration), and nPVI-C (durational differences between adjacent consonantal intervals), as indicated in a MANOVA with fixed factor Level (A1–C2), which showed a significant main effect for %V, $F(5, 141) = 2.33$, $p < 0.05$; Varco-C, $F(5, 141) = 4.02$, $p < 0.01$; and nPVI-C, $F(2, 148) = 2.65$, $p < 0.05$. Post hoc tests (Scheffe) showed that for Varco-C, the effect was attributable to a marginal difference between A1 and A2 ($p = 0.078$), between A2 and C1 ($p < 0.05$), and A2 and C2 ($p < 0.05$); and for nPVI-C, A2 and C2 differed ($p < 0.05$); for %V none of the individual comparisons reached significance.

Overall, the consonantal metrics showed a steady increase over proficiency levels, with the exception of Level A1. This general upward trend indicates that the variability in consonantal interval durations as well as the durational differences between adjacent consonantal intervals increased, suggesting that as learners improved in proficiency, they became more adept at dealing with both single consonants and more complex consonant clusters, and showed a more or less steady progression in consonantal variability. Interestingly, the Varco-C measure was found to be stable at the higher C1/C2 levels, while nPVI-C increased at those levels (but not significantly). This indicates that C level L2 speakers are appropriately varying the durations of consonantal sequences in their speech globally (i.e. across the utterance).

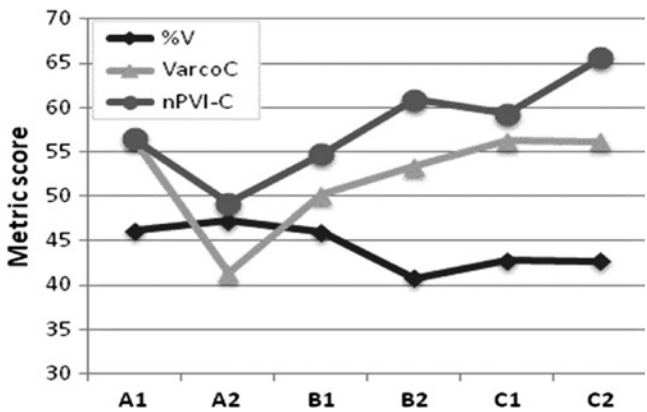


Figure 9.3 Mean metric values for %V, VarcoC and nPVI-C at 6 CEFR levels

This is likely to primarily reflect the adjustments that need to be made when consonantal sequences vary in complexity – and results in the same rhythmic profile at the two highest levels in Varco-C. However, when the variation in the duration of consonantal intervals is considered more locally (i.e. in terms of pairwise variability between consonants), the values are still changing, which could suggest that more localized variation in duration due to factors like accentuation and boundary marking is still developing further at this level. In the case of the vocalic metric %V (the proportion of vocalic material in speech), a shift occurred between B1 and B2, indicating that as learners developed in proficiency between these two levels, they became more adept at producing reduced vowels, and at decreasing the number of (incorrectly inserted) vowels.

Moving on to the prosodic lengthening measures, the durations of syllables overall and by syllable type were investigated. As a starting point, the mean number of syllables produced by learners at each level was calculated (Figure 9.4). Even though this measure is not strictly speaking an acoustic-phonetic measure, and is typically considered a measure of fluency, it is a useful initial gauge of the development of learners across levels. Findings indicated a clear progression of mean number of syllables across proficiency levels, with the exception of Levels A1 and A2. As the learners in the sample developed in proficiency, they produced more syllables in a 60-second stretch.

The duration of syllables – which reflects learners' speech planning and execution processes (Field, 2011; Levelt, 1989) – was also examined. Logically, if more syllables are produced in a constant stretch of speech (e.g. 60 seconds),

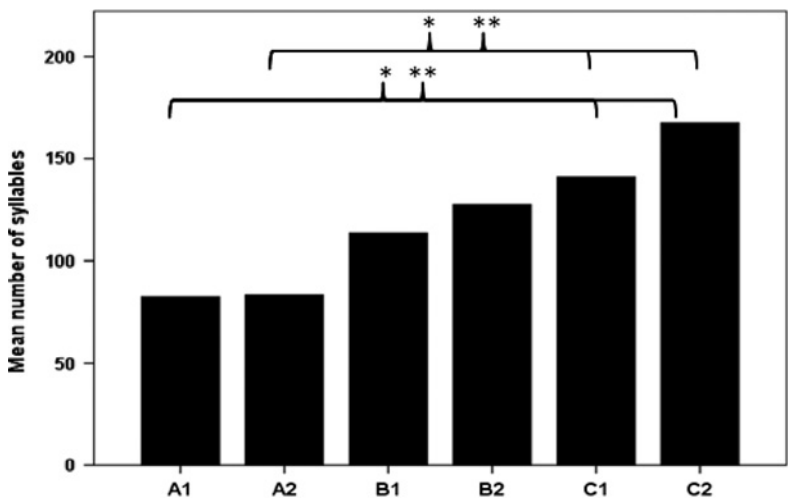


Figure 9.4 Speaking rate: Mean number of syllables (in 60-second sample)
Note: *$p < 0.05$; ** $p < 0.01$.

then the syllables will become shorter. In less proficient learners, it is expected that the speech planning process will be slower as a result of lack of automatization of cognitive processes and limited lexico-grammatical resources (Levelt, 1989). The execution is further slowed down by the need for speech articulators to be moved into positions that could be unfamiliar for learners (e.g. 'th' or realizations of /r/ in different languages), which takes a longer time. As Levelt (1989: 413) noted, fluent articulation 'involves the co-ordinated use of approximately 100 muscles' – a tall order for a language learner.

Conforming to expectations, the analyses revealed that syllable durations in less proficient learners are significantly longer and decrease with increasing proficiency, as seen in the downward trend in Figure 9.5. An ANOVA of mean syllable duration showed a significant effect for Level, $F(5, 1436) = 37.099$, $p < 0.01$. Post hoc tests (Scheffe) confirmed the three-way grouping that is visible in the figure (Levels A1/A2, B1/B2, C1/C2).

The specific durations of different prosodic syllable types (i.e. unaccented, accented, unaccented final and accented final) were additionally examined as a potential discriminating measure across proficiency levels. An ANOVA with CEFR Level (six levels) and Prosodic position (four levels) as fixed factors showed significant main effects for both, $F(5, 1418) = 10.93$, $p < 0.001$ and $F(3, 1418) = 146.09$, $p < 0.001$, respectively, as well as a two-way interaction, $F(15, 1418) = 2.68$, $p < 0.001$. Post hoc tests (Scheffe) showed that accented and unaccented final syllables generally formed three

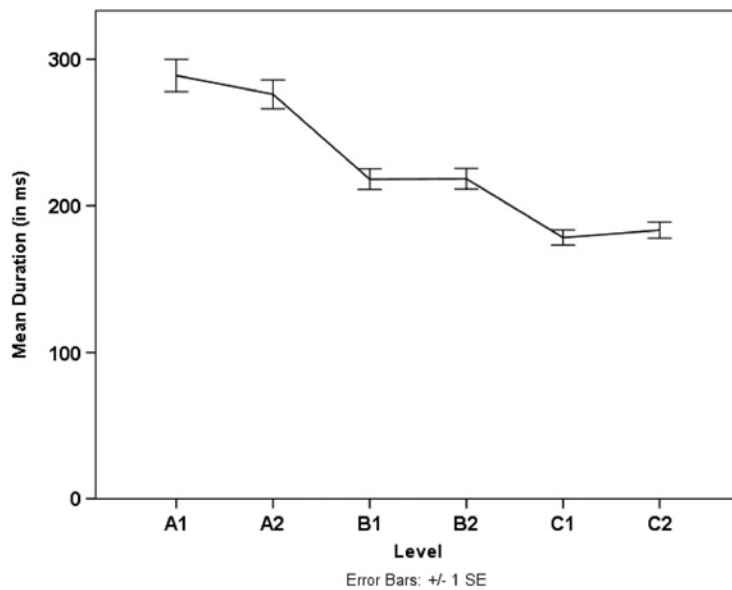


Figure 9.5 Speaking rate: Mean duration of syllables

homogeneous subsets: Levels A1/A2, B1/B2 and C1/C2, as the measurements in Figure 9.6 indicate.

Figure 9.6 indicates that, with the exception of unaccented final syllables, the duration of all other syllable types generally decreases with increasing proficiency, which is a reflection of developing speech planning and execution processes. However, importantly, the durations of the different syllable types do not change to the same extent. Instead, unaccented and accented durations differ more at advanced C levels than at beginner A levels. This suggests a progression towards the more English realization of the syllable types where consistent durational differences are present between the two syllable types. Additionally, the lengthening of unaccented final syllables, a very characteristic property of stress-timed languages such as English, becomes longer than unaccented syllables at the C levels but is indistinguishable from accented syllables at the A levels. Therefore, while the durations of syllable types clearly overlap at beginner levels and are thus indistinguishable, at C2 learners have implemented different 'categories' for all prosodic syllable types that are marked by distinct durational patterns. This shows a progression from a more syllable-timed realization of syllable types to a more stress-timed English realization.

It is also worth noting the error bars in Figure 9.6, which indicate variability within proficiency levels and overlap between scale bands – a finding

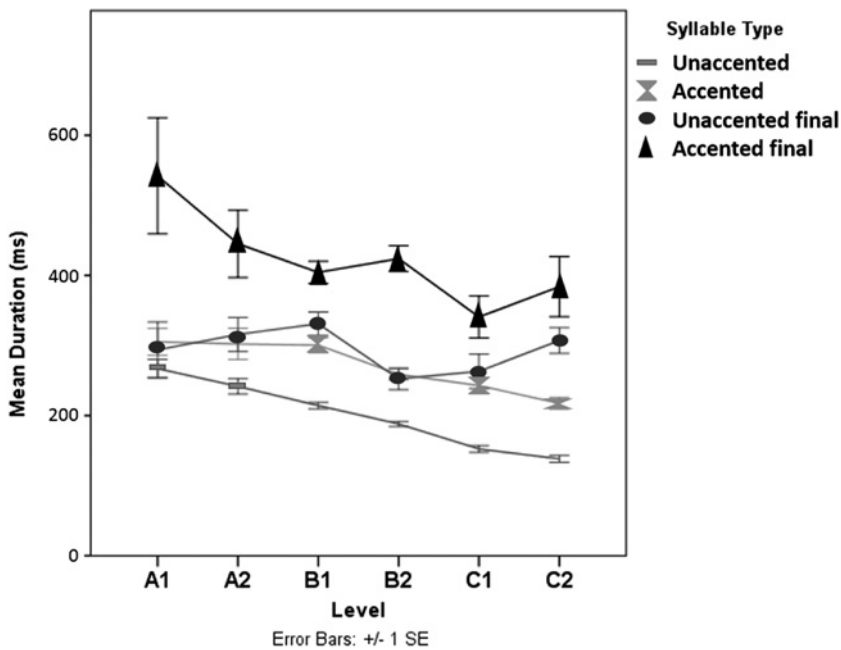


Figure 9.6 Mean prosodic lengthening across proficiency levels by syllable type

that echoes results reported in Kang (2013) and Iwashita *et al.* (2008). Since this study is an exploratory study with only two speakers per CEFR level, more research is needed to further validate the robustness of the effects between different proficiency levels. Nevertheless, we can conclude that the measures tested here are useful in discriminating between levels of proficiency, as demonstrated with the Spanish learners of English participating in the study.

Study 2: Differences across first languages

Figure 9.7 plots the proportion of vocalic material (%V) produced by the German, Spanish and Korean learners against the variability in their consonantal intervals (Varco-C); rPVI-C is omitted from the figure, since its pattern of results resembles that for Varco-C. The data in the figure show that the rhythm metrics differ for the three L1 learner groups, with the highest Varco-C values for German and the highest %V values for Spanish, as would be expected under L1 transfer – that is, a high number of consonants and low number of vowels in German as a stress-timed language and vice versa in Spanish as a syllable-timed language.

A MANOVA, which included all six rhythm metrics and the fixed factors Language (German, Spanish, Korean) and Level (B1, B2), revealed a significant main effect of Language for Varco-C, $F(2, 148) = 3.36$, $p < 0.05$, and rPVI-C, $F(2, 148) = 4.70$, $p = 0.01$, and a significant interaction between Language and Level for %V, $F(2, 148) = 3.45$, $p < 0.05$, and rPVI-C, $F(2, 148) = 3.28$, $p < 0.05$, but no main effect for Level. Post hoc tests (Scheffe)

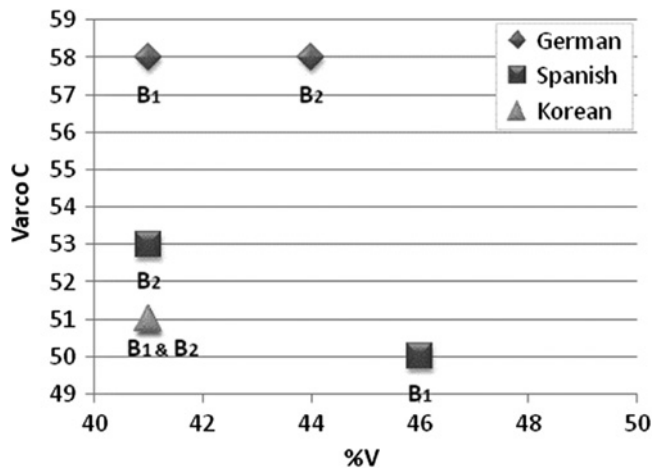


Figure 9.7 Mean metric values for %V and Varco-C for German, Spanish and Korean learners of English at Levels B1 and B2

showed that German and Korean learners differed marginally for Varco-C ($p = 0.071$), and German and Spanish learners did so for rPVI-C ($p < 0.05$).

Figures 9.8 and 9.9 show the effect of L1 background on accentual and phrase-final lengthening for Levels B1 and B2, respectively. Figure 9.8 reveals an L1 background effect, with German learners better at distinguishing between the prosodic lengthening of accented and unaccented final syllables than the other learners, at least at Level B1. At Level B2 (Figure 9.9) the picture changes, with German learners differentiating less between the different syllable types, while the Koreans are doing better, and the Spanish are

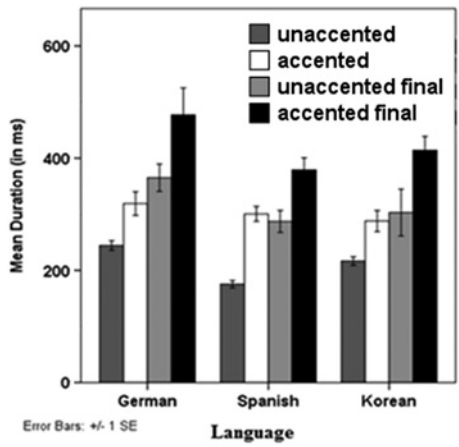


Figure 9.8 Lengthening across languages by syllable type (Level B1)

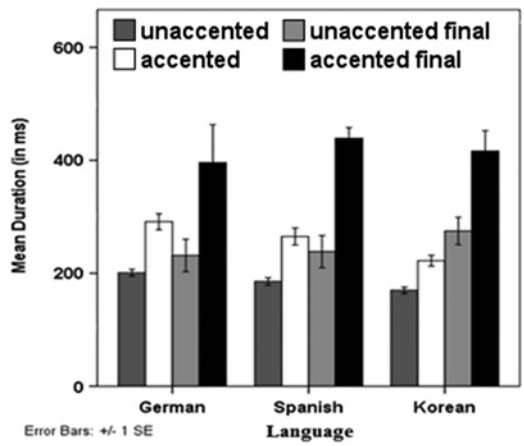


Figure 9.9 Lengthening across languages by syllable type (Level B2)

essentially continuing as before. Interestingly, although German learners with the most typologically similar language background to the target language showed an early advantage for some measures, they were often still as far off target as the other L1 groups at the intermediate B levels.

An ANOVA with fixed factors L1 Language (German, Spanish and Korean), Level (B1 and B2), and Lengthening condition (unaccented, accented, unaccented final, accented final) showed significant main effects for all three factors respectively, $F(2, 1683) = 3.02$, $p < 0.05$; $F(1, 1683) = 11.99$, $p = 0.001$; $F(3, 1683) = 116.84$, $p < 0.001$, and a significant interaction between Language and Level, $F(2, 1683) = 3.22$, $p < 0.05$, confirming the pattern of results sketched above. Therefore, we can conclude that the measures tested here can successfully distinguish between the spoken productions of learners from different L1 backgrounds.

Discussion

Overall, this investigation has shown that some of the prosodic measures under investigation here can provide useful micro-level prosodic measures for consideration in L2 teaching and assessment contexts. The results of Study 1, which focused on differences across proficiency levels, indicated that the learner speech observed at the different proficiency levels signalled progression from a more syllable-timed realization of speech to a more stress-timed realization, in line with the prosodic requirements of English. More specifically, the higher level learners in the sample used here were found to:

(1) have a higher speech rate, and produce more frequent and shorter syllables as reflected in the speech rate measurements, likely as a result of higher automaticity of speech planning and execution processes (Field, 2011);
(2) differ durationally between unaccented/accented and non-final/final syllables, indicating that at higher levels of proficiency learners are likely to have implemented different prosodic categories for different syllable types and display these with appropriate durational patterns;
(3) be more adept at producing appropriate durations for vowels and consonants as seen in the %V metric which showed a downward shift between Levels B1 and B2, and in the Varco-C and nPVI-C metrics which showed a more-or-less steady increase in consonantal variability from A2 to C2; this most likely reflects increased mastery of language-specific properties like vowel reduction and syllable structure complexity, and also accentual and final lengthening.

These findings are in line with Low's (2006) assertion that the stress-timing rhythm patterns of English speech would present problems for

learners. They also support previous research which has shown an increase in syllables in learner speech as proficiency increases (Kang, 2013; Kang & Wang, 2014) and the difficulty of lower level learners with managing stress-timed speech (Iwashita *et al.*, 2008). The current findings extend earlier research through pinpointing micro-level features of rhythm, which cause challenges for learners.

The results of Study 2, which focused on learners from three typologically different L1 backgrounds at intermediate-level proficiency, indicated that learners with different language backgrounds show different prosodic patterns:

(1) learners from a stress-timed language such as German showed the highest ability to deal with consonantal variation (as seen in Varco-C and rPVI-C), and learners from a syllable-timed language such as Spanish displayed the highest proportion of vocalic material (%V), suggesting that transfer effects that can be predicted on the basis of the rhythmic properties of their respective L1s still take effect at the B levels;
(2) German learners were better at distinguishing the durations between different syllable types than the Korean and Spanish learners at B1, but the Korean learners showed a clear progression at B2.

These findings suggest that there may be a prosodic basis to the results reported by Crowther *et al.* (2015), who found that speakers' L1 plays an important role in listener judgements of L2 comprehensibility.

A finding common to both studies was the variability within CEFR levels, with differences found mainly between Levels A1 and C2, and no meaningful differences between adjacent levels. This is in line with results reported in other studies, which have found that pronunciation features are not clearly distinct between adjacent levels (Isaacs *et al.*, 2015; Iwashita *et al.*, 2008; Kang, 2013; Kang & Wang, 2014).

Implications

Although the implications of these findings are only tentative due to the small-scale nature of the study, there are nevertheless useful insights for learning, teaching and assessment. In L2 pronunciation, global intelligibility and local (phonetic) precision are key concerns, and teachers and examiners must, therefore, have knowledge of the components of pronunciation and an awareness of when, how and why a pronunciation feature is problematic for learners. Levis (2006: 247) rightly argues that 'the first thing that teachers must learn is to give more than global impressions of pronunciation. They need to become aware of relevant phonological categories and be able to name important errors.' What this study has revealed are the rhythmic

challenges that learners might face, such as distinguishing between different syllable types in terms of duration, the difficulty of reducing the amount of vocalic material in their speech, and the challenge of varying consonantal intervals. Important prosodic features such as syllable variability and the crucial role of vowel reduction could, as such, become a focus in the classroom, as also suggested by Liang (2003) and Low (2006). Such awareness would allow teachers to include prosodic features in their instruction, and would assist them to focus on the form-meaning relationships which are fundamental to learning (Isaacs, 2009). Understanding what learners can and cannot do at a specific proficiency level could also support learners to notice key aspects of their speech that would aid learning through conscious noticing and awareness (Schmidt, 1990). Clarity of pronunciation features that matter and affect their speech can additionally help learners to become more autonomous. The findings also indicate that different pronunciation features could be fruitfully emphasized in teaching and learning based on the L1 background of learners.

Investigating measures which 'count' in pronunciation also has implications for the assessment of pronunciation. As noted at the beginning of this chapter, raters have reported lower confidence in assessing pronunciation as compared to other skills (e.g. Brown & Taylor, 2006) and are more comfortable making global judgements of intelligibility as opposed to nuanced judgements about features of pronunciation. These findings may reflect the 'undemanding nature of judgements [of global intelligibility] in terms of technical expertise' (Yates *et al.*, 2011: 36), as opposed to the precise technical knowledge needed to deconstruct pronunciation into its constituent parts, and they signal the need for a comprehensive and in-depth training of examiners. The role of L1 background in learner speech has also confirmed the importance of supporting examiners in developing an 'international ear' through exposure to test takers from different L1s and levels of intelligibility, since a German and a Spanish learner, for example, would present different profiles that examiners would need to evaluate – as underlined by our findings. Raters' increased familiarity with L2 speech by learners from different L1 backgrounds and ability levels would, in addition, minimize the effect that rater familiarity with a test taker's pronunciation plays in assessment – empirically shown to play a role (e.g. Ballard & Winke, this volume; Carey *et al*., 2011; Saito *et al*., this volume) – and would ultimately result in more reliable and valid assessment. The role of L1 background in learner speech and its possible effect on rater comprehension of that speech also indirectly lends support to Wagner and Toth's (this volume) argument that pronunciation needs to be considered as part of both the speaking and listening constructs.

The findings have further implications for assessment in terms of scale development, scale descriptors and number of scale bands. Regarding the number of bands, the findings reported here indicate that with some of the

measures there were three distinct homogenous subsets in the data – A1/A2, B1/B2, C1/C2, suggesting that at least in the small-scale study here, only three broad levels were observed. This finding is, of course, based on a small subset of measures for pronunciation, but it provides some confirmation of research that has reported difficulty in scaling pronunciation ability across six or more levels. For example, in an older version of the IELTS test, pronunciation was marked on a four-point scale, in contrast to the nine-point scale used for other traits such as lexical and grammatical resource, coherence and fluency, largely as a result of the different Many-Facet Rasch analysis findings for pronunciation (Taylor, personal communication). Research on the current nine-point IELTS scale – even though it was reported by examiners to be easy to use – has, nevertheless, indicated that listener-rated measures were difficult to discriminate between nine levels (Isaacs *et al.*, 2015). This finding also echoes the difficulties reported by North and Hughes (2003) in developing a six-level pronunciation scale, and indicates that pronunciation ability may be more meaningfully measured with fewer scale levels, or at least that further empirical work is needed focusing on the scaling of pronunciation.

The micro-level features identified here could also impact on the development of descriptors in assessment scales. They indicate that suprasegmental features, and specifically rhythm, play a distinguishing role at both higher and lower levels of L2 proficiency, and should therefore be included at all levels (as also argued by Harding, this volume). Not all features identified as important in this study can be captured in assessment scales, since scales are driven by a need for conciseness, as well as a need for usability and the avoidance of vague terminology, as Harding contends in an earlier chapter in this volume. This presents a case for more explicit references in scales to components of rhythm. Examples of this approach can be seen in the IELTS scale, which includes descriptors such as 'can sustain appropriate rhythm' and 'rhythm may be affected by lack of stress-timing' (http://www.ielts.org). Such specific references to the components of prosody in assessment scales are positive examples of how pronunciation can be deconstructed in scales, and how it could potentially make examiners more reliable through training and awareness raising of important pronunciation features. For example, in a survey of examiner views on the IELTS speaking scales, 83% and 76% of examiners reported that 'rhythm' and 'stress timing', respectively, were salient pronunciation features for them when they assess (Galaczi *et al.*, 2012).

Even if not included in assessment scales, micro-level features can provide useful guidance for examiner training, since they provide explicit information about what matters in pronunciation across CEFR levels, and can be beneficial in developing a shared understanding of pronunciation and its constituent parts. As Yates *et al.* (2011: 36) argue, identification of important micro-level features could offer examiners 'a discourse that they can use to

articulate what they have noticed' and 'a framework within which to talk about the same aspects of a performance'.

The findings reported here could also have implications for automated speech recognition (ASR) and assessment systems of speech in which pronunciation measurement plays a central role. Most current systems are based on detailed taxonomies of pronunciation features (e.g. articulation and duration of phonemes, pauses, use of pitch, and mean duration between stressed syllables) and associated weightings of those features in computer algorithms (e.g. Evanini & Wang, 2013; van Moere, 2012; Xi *et al.*, 2012). Such systems could potentially include some of the rhythm measures identified here as a means of improving speech recognition accuracy and enhancing the assessment of prosody. The findings on the effect of L1 background on rhythm can also be useful for ASR, as they support the development of ASR systems targeted at specific L1s or language groups.

Future Research and Conclusion

This exploratory study is constrained by its small-scale nature and limited generalizability. A larger scale investigation would address this limitation and be useful in further exploring the pronunciation measures that were found to play a role. Such an investigation could extend not just to a larger sample of learners and L1s, but also to task types, since research has revealed interesting findings about differences in pronunciation measures across monologic and interactional task types (Kang & Wang, 2014). A mixed-methods integration of quantitative and qualitative findings would provide further useful insights, and could, for example, extend the present study to an investigation of the degree to which raters attend to the instrumentally derived measures identified here. The need for longitudinal work also needs to be borne in mind, since it could complement the cross-sectional snapshots provided in the current study. The use of data from two different assessment modes (computer and face-to-face) – a potential limitation of the study – is a further area to explore. Even though from an assessment perspective there are clear differences between the two test modes, this is not considered to be a major threat to this study, since in both modes learners provided extemporaneous speech which illustrated their mastery of a range of phonological features. Nevertheless, an exploration of the effect of the face-to-face versus computer-based mode on pronunciation could reveal potential pronunciation differences and implications for assessment. Notwithstanding these limitations, a systematic empirical investigation of the research questions guiding this study has contributed to an empirically based understanding of the pronunciation construct, which can inform scale development and rater and teacher competence, and contribute to the development and assessment of learner pronunciation.

References

Abercrombie, D. (1967) *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

ACTFL (American Council on the Teaching of Foreign Languages) (2012) *ACTFL Proficiency Guidelines 2012*. Alexandria, VA: American Council on the Teaching of Foreign Languages.

Anderson-Hsieh, J., Johnson, R. and Koehler, K. (1992) The relationship between native speaker judgements of nonnative pronunciation and deviance in segmentals, prosody and syllable structure. *Language Learning* 42, 529–555.

Bachman, L. (1990) *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Boersma, P. and Weenink, D. (2010) *Praat: Doing Phonetics by Computer*, Version 5.2 [Computer program].

Bolinger, D.L. (1965) Pitch accent and sentence rhythm. In I. Abe and T. Kanekiyo (eds) *Forms of English: Accent, Morpheme, Order* (pp. 139–180). Cambridge, MA: Harvard University Press.

Brown, A. and Taylor, L. (2006) A worldwide survey of examiners' views and experience of the revised IELTS speaking test. *Research Notes* 26, 14–18.

Carey, M.D., Mannell, R.H. and Dunn, P.K. (2011) Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing* 28, 201–219.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Crowther, D., Trofimovich, P., Saito, K. and Isaacs, T. (2015) Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly* 49, 814–837.

Dauer, R.M. (1983) Stress-timing and syllable-timing reanalysed. *Journal of Phonetics* 11, 51–62.

Dellwo, V. (2006) Rhythm and speech rate: A variation coefficient for ΔC. In P. Karnowski and I. Szigeti (eds) *Language and Language Processing: Proceedings of the 38th Linguistic Colloquium, Piliscsaba 2003* (pp. 231–241). Frankfurt am Main: Peter Lang.

Evanini, K. and Wang, X. (2013) Automated speech scoring for non-native middle school students with multiple task types. *Proceedings of INTERSPEECH, Lyon, France, 2013*. Baixas: ISCA.

Field, J. (2005) Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly* 39, 399–424.

Field, J. (2011) Cognitive validity. In L. Taylor (ed.) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*. Studies in Language Testing (Vol. 30, pp. 65–111). Cambridge: UCLES/Cambridge University Press.

Galaczi, E.D., Lim, G. and Khabbazbashi, N. (2012) Descriptor salience and clarity in rating scale development and evaluation. Paper presented at the Language Testing Forum, Bristol.

Grabe, E. and Low, E.L. (2002) Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven and N. Warner (eds) *Laboratory Phonology 7* (pp. 515–546). Berlin: Mouton de Gruyter.

Guilbault, C.P.G. (2002) The acquisition of French rhythm by English second language learners. Doctoral dissertation, University of Alberta.

Gussenhoven, C. and Jacobs, H. (2005) *Understanding Phonology* (2nd edn). London: Hodder Education.

Gut, U. (2009) *Non-native Speech: A Corpus-based Analysis of Phonological and Phonetic Properties of L2 English and German*. Frankfurt am Main/Oxford: Peter Lang.

Gutiérrez-Díaz, F. (2001) The acquisition of English syllable timing by native Spanish learners of English. An empirical study. *International Journal of English Studies* 1 (1), 93–113.
Hahn, L.D. (2004) Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly* 38, 201–223.
Isaacs, T. (2009) Integrating form and meaning in L2 pronunciation instruction. *TESL Canada Journal* 27, 1–12.
Isaacs, T. (2014) Assessing pronunciation. In A.J. Kunnan (ed.) *The Companion to Language Assessment*. Hoboken, NJ: John Wiley.
Isaacs, T. and Trofimovich, P. (2012) 'Deconstructing' comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition* 34, 475–505.
Isaacs, T., Trofimovich, P., Yu, G. and Chereau, B.M. (2015) Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS Research Reports Online* 4, 1–48.
Iwashita, N., Brown, A., Mcnamara, T. and O'Hagan, S. (2008) Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics* 29 (1), 24–49.
Jenkins, J. (2002) A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics* 23, 83–103.
Kang, O. (2010) Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System* 38 (2), 301–315.
Kang, O. (2013) Linguistic analysis of speaking features distinguishing general English exams at CEFR levels. *Research Notes* 52, 40–48.
Kang, O. and Wang, L. (2014) Impact of different task types on candidates' speaking performances and interactive features that distinguish between CEFR levels. *Research Notes* 57, 40–49.
Kang, O., Rubin, D.L. and Pickering, L. (2010) Suprasegmental measures of accentedness and judgements of English language learner proficiency in oral English. *Modern Language Journal* 94, 554–566.
Kohler, K. (1982) Rhythmus im Deutschen [Rhythm in German]. *Arbeitsberichte, Institut für Phonetik der Universität Kiel* 19, 89–106.
Levelt, W.J.M. (1989) *Speaking*. Cambridge, MA: MIT Press.
Levis, J.M. (2006) Pronunciation and the assessment of spoken language. In R. Hughes (ed.) *Spoken English, TESOL and Applied Linguistics* (pp. 245–270). New York: Palgrave Macmillan.
Li, A. and Post, B. (2014) L2 acquisition of prosodic properties of speech rhythm: Evidence from L1 Mandarin and German learners of English. *Studies in Second Language Acquisition* 36 (2), 223–255.
Liang, W.X. (2003) Teaching weak forms. *Forum* 41, 32–36.
Linacre, J. (1989) *Many-Facet Rasch Measurement*. Chicago, IL: MESA Press.
Low, E.L. (1998) Prosodic prominence in Singapore English. Unpublished doctoral dissertation, University of Cambridge.
Low, E.L. (2006) A review of recent research on speech rhythm: Some insights for language acquisition, language disorders and language learning. In R. Hughes (ed.) *Spoken English, TESOL and Applied Linguistics* (pp. 99–125). New York: Palgrave Macmillan.
Low, E.L., Grabe, E. and Nolan, F. (2000) Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech* 43 (4), 377–401.
Major, R. (2008) Transfer in second language phonology: A review. In J.G.H. Edwards and M.L. Zampini (eds) *Phonology and Second Language Acquisition* (pp. 63–94). Amsterdam: John Benjamins.
Munro, M.J. and Derwing, T.M. (1995) Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning* 45, 73–97.

North, B. and Hughes, G. (2003) CEF illustrative performance samples. See www.eaquals.org/wp-content/uploads/Documentation-to-English-2004-DVD-with-adults_speaking.pdf.

Peterson, G.E. and Lehiste, I. (1960) Duration of syllable nuclei in English. *Journal of the Acoustical Society of America* 32 (6), 693–703.

Pickering, L. (2001) The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly* 35 (2), 233–255.

Pike, K.L. (1945) *The Intonation of American English*. Ann Arbor, MI: University of Michigan.

Post, B., Payne, E., Prieto, P., Astruc, L. and Vanrell, M.D. (2010) A multisystemic model of rhythm development: Phonological and prosodic factors. BAAP Colloquium, London, 29–31 March.

Prieto, P., Vanrell, M.d.M., Astruc, L., Payne, E. and Post, B. (2012) Phonotactic and phrasal properties of speech rhythm: Evidence from Catalan, English, and Spanish. *Speech Communication* 54, 681–702.

Pulgram, E. (1970) *Syllable, Word, Nexus, Cursus*. The Hague: Mouton.

Ramus, F., Nespor, M. and Mehler, J. (1999) Correlates of linguistic rhythm in the speech signal. *Cognition* 73, 265–292.

Ramus, F., Dupoux, E. and Mehler, J. (2003) The psychological reality of rhythm classes: Perceptual studies. In M. Solé, D. Recasens and J. Romero (eds) *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 337–342). Barcelona: Universitat Autonomá de Barcelona.

Roach, P. (1982) On the distinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal (ed.) *Linguistic Controversies* (pp. 73–79). London: Edward Arnold.

Schmidt, R. (1990) The role of consciousness in second language learning. *Applied Linguistics* 11, 129–150.

Seong, C. (1995) The experimental phonetic study of standard current Korean speech rhythm: With respect to its temporal structure. PhD thesis, Seoul National University.

Trofimovich, P. and Baker, W. (2006) Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition* 28, 1–30.

Trofimovich, P. and Isaacs, T. (2012) Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition* 15 (4), 905–916.

Van Moere, A. (2012) A psycholinguistic approach to oral language assessment. *Language Testing* 29 (3), 325–344.

White, L. and Mattys, S.L. (2007) Calibrating rhythm: First language and second language studies. *Journal of Phonetics* 35, 501–522.

Xi, X., Higgins, D., Zechner, K. and Williamson, D. (2012) A comparison of two scoring models for an automated speech scoring system. *Language Testing* 29 (3), 371–394.

Yates, L., Zielinski, B. and Pryor, E. (2011) The assessment of pronunciation and the new IELTS Pronunciation Scale. In J. Osborne (ed.) *IELTS Research Reports* 12 (pp. 1–46). Melbourne and Manchester: IDP IELTS Australia and British Council.

Part 4

Sociolinguistic, Cross-cultural and Lingua Franca Perspectives in Pronunciation Assessment

10 Commentary on the Native Speaker Status in Pronunciation Research[1]

Alan Davies

L'accent du pays où l'on est né demeure dans l'esprit et dans le coeur comme dans le langage.
[The accent of one's birthplace remains as much in one's spirit and heart as in one's speech.]
Duc de la Rochefoucauld, 1613–1680: Maximes, 1678, No. 342 (1822)

There is no easy way of defining the native speaker for, as Hyltenstam and Abrahamsson (2000) point out, the term can be defined in a number of ways, each of which is defensible. They include: (a) native speaker by birth (that is, by early childhood exposure); (b) native speaker by virtue of being a native user; (c) native speaker (or native speaker like) by being an exceptional learner; (d) native speaker through education in the target language medium; and (e) native speaker through long residence in the adopted country.

There are further problems: the English article *the* is inappropriate (*the native speaker*). There are native speakers, but *the* native speaker does not exist. Each of us is a native speaker of something, some code, dialect or language, but each of us is different. Even siblings brought up together in the same home are not identical, which explains why it is that we can place which brother or sister is speaking on the telephone, even before they have said who they are. To say that we all differ from one another does not, of course, mean that there is no sharing. There is more in common among those who regard themselves as native speakers of language X than there is between them and native speakers of language Y (de Saussure, 1916). For the time being, I ignore those distant dialect speakers of X whose variety takes them to the border where it overlaps with distant dialect speakers of Y.

And yet, the native speaker is both contentious and necessary. The native speaker is contentious as the critical attacks against the power of the native speaker attest (Braine, 1999; Canagarajah, 1999; Cook, 1999; Holliday, 2005; Medgyes, 1992). What is less often noted is that these attacks are almost all

against the native speaker of English, which suggests that the issue is more political than linguistic, postcolonial, even racist in a world currently dominated by the necessity of English. French, similarly located in a postcolonial critique, does not attract such reproach, or rather the critique here is for the loss of négritude (i.e. African francophone) identity (Davies, 2013).

The native speaker is necessary for the same reason; international communication requires English, which in turn means very large scale programmes in the teaching and learning of English as a foreign language (EFL). For such programmes to succeed, what is essential is a described and assessable model for the curriculum, for the textbook, for the examination – a need which support for English as a lingua franca (Davies, 2013; Elder & Davies, 2006; Jenkins, 2005; Seidlhofer, 2004) has not come to terms with.

The native speaker is attacked even more widely. The American Charles Ferguson, first Director of the Center for Applied Linguistics in Washington, DC, wrote:

> Linguists … have long given a special place to the native speaker as the only true and reliable source of language data … much of the world's verbal communication takes place by means of languages which are not the users' mother tongue, but their second, third or nth language, acquired one way or another and used when appropriate. This kind of language use merits the attention of linguists as much as do the more traditional objects of their research … the whole mystique of native speaker and mother tongue should preferably be quietly dropped from the linguists' set of professional myths about language. (Ferguson, 1983: vii)

And Noam Chomsky goes even further:

> the question of what are the 'languages' or 'dialects' attained and what is the difference between 'native' and 'non-native' is just pointless. (Chomsky, 1985)

But if there is no agreed definition, no describable content for a teaching programme posited on the native speaker; what model can take its place? That is, if the native speaker were dropped, what could serve in its stead? The answer to this dilemma to which Chomsky refers is that there is a model which is, in fact, in common use, and that model is the *idealized native speaker* (i.e. not a real entity but an aspirational construct). Does this bring us any closer to a solution for our model? The answer is yes, because in the absence of an adequate description of the native speaker, what takes its place is the Standard Language.

Such a solution may appear glib in spite of the lack of boundaries of the English Standard Language (i.e. French, like other languages with institutionalized Academies, may lend itself more readily to a description

of its Standard Language than does English). Yet the Standard Language model (=idealized native speaker) provides an agreed-upon goal that is applicable worldwide.

But problems remain. In particular, the domain of the Standard Language is the written language (Halliday, 2003). What this implies is that agreement on the Standard Language goal in reading and writing does not require equal agreement on the way that language is or should be spoken. At its extreme, what this means is that while the same content is taught in classrooms worldwide, that content is only for the written language and therefore teachers and students may well speak to one another in the classroom in their local language. Closer to home it means that wide regional variation in speech (Scottish, South English, etc.) is acceptable, but no variation in the grammar and styles of the written language is similarly acceptable. Even vocabulary differences between Australian and British speakers in their spoken language, for example, are unlikely to find their way into the written English of educated Australians and British.

Does this mean that the spoken language has no standard variety? The answer has to be a qualified no. Since language is so potent a marker of status and class, it is not surprising that there is a tendency for speakers both in their first language (L1) and their second language (L2) to accommodate towards the prestige accent, or indeed to a prestige accent where there is more than one (Davies, 2013). Thus, in the inner circle of English-using nations (Kachru, 1992), there is one or more prestigious way of speaking by the educated. In other settings (the outer circle; the expanding circle), it is likely that the unstated model, which not all achieve, will be a British modified Received Pronunciation (RP) or an American educated accent.

Standard language in itself does not have an accent, but there are more and less prestige accents. It is probably the case that there is always one or more prestige accent. For British English it is RP, which is the accent associated with the independent (i.e. private, fee-charging) schools in the UK, the BBC, the royal family and so forth. But even in situations where there is no institutionalized standard language, there is still likely to be a prestige model of spoken delivery. Leonard Bloomfield, the American linguist who studied the Native American Menomini community, comments that the Menomini 'will say that one person speaks well and another badly, that such-and-such a form of speech is incorrect and sounds bad, and another too much like a shaman's preaching or archaic ("the way the old, old people talked")' (Bloomfield, 1927: 89). Bloomfield notes that although a foreigner, he was able to share in these judgements of the Menomini:

> The nearest approach to an explanation of good and bad language seems to be this, then, that by a cumulation of obvious superiorities, both of character and standing as well as of language, some persons are felt to be better models of conduct and speech than others. (Bloomfield, 1927: 93)

The issue of accent prestige has been much discussed with particular reference to the L1 (e.g. Wolfram, 2004). Sociolinguists researching variation, such as Mugglestone (1995), tend to view prestige as an irrelevant variable, while applied linguists such as Honey (1989) recognize that in the use of language in the community, prestige is crucial. Honey has been unfairly criticized for taking this position, which is said to be supportive of class divide. However, he is in fact pointing out just what the situation is, not that he approves of it. In other words, there is always a socially accepted prestige model of accent.

> Perhaps because accent is more resistant to change than dialect (hence foreign accents) and more easily identified with origin and identity, there is little emphasis today on using education to change accent. Even so, it does seem that one of the effects of education (however indirect) is to bring about some accommodation towards a norm with prestige (Davies, 2005: 4–5).

Such attempts at accommodation are usually examined with regard to a non-prestige L1 accent. But it also applies to foreign accents. After all, just as an L1 accent defines one's provenance, so too does an L2 foreign accent, which is often shared with family and peers and even larger social and regional groups. Thus, it is possible for most of us to label a speaker of English with a foreign accent as, say, French; it is also possible for a phonologist to point to a much narrower background, say, Provencal or Normandy French. So what is accent? David Crystal defines accent thus:

> The cumulative auditory effect of those features of pronunciation which identify where a person is from, regionally or socially … the term refers to pronunciation only, it is thus distinct from dialect, which refers to grammar and vocabulary as well. … In Britain, the best example is the regionally neutral accent associated with a public school education, and of the related professional domains, such as the Civil Service, the law courts, the Court and the BBC: hence the labels Queen's English, BBC English, and the like. Received Pronunciation (RP) is the name given to this accent and because of its regional neutrality, RP speakers are sometimes thought of as having no accent. This is a misleading way of putting it, however: linguistics stresses that everyone must have an accent, though it may not indicate regional origin. (Crystal, 1997: 2)

Many years ago, I was invited by the West African Examinations Council to advise on the introduction of a compulsory spoken English test at school certificate level, effectively the final examination at secondary school and

the examination that, at that time, determined entry to university. There was a test already in use, the McCallien Test, but this was regarded as in need of renewal. It tested only RP segmental sounds and was too difficult for the majority of students who had no contact with RP. McCallien herself had an educated Scottish English accent, certainly not RP; the teachers of those presenting for the examination spoke with some form of West African accent. And so, the assumed goal of the test was not available to most students. The McCallien Test also left out all other aspects of spoken communication. We proposed a new test which used as a norm an educated West African accent. But there was little appetite for such a change: stakeholders were worried that to promulgate one or more West African norms might ghettoize West African speakers of English, who would be labelled as second-class English speakers. The initiative collapsed; for me the experience illustrated the difficulty of change in norm setting.

In her discussion of foreign language accent, Moyer (2013) accepts the need for a holistic approach to the spoken language and takes us through the various aspects that contribute to the formation of a foreign language accent and to individuals' uncertainty as to whether to aim at sounding more or less nativelike. Individuals differ markedly in how far they are prepared to accommodate across accent boundaries (e.g. L1: Scottish and southeast England English; L2: French-English). Foreign accent, Moyer helpfully points out, is not confined to segmental distinctions:

> Intonation, loudness, pitch, rhythm, length, juncture and stress are among accent's many features; all of which classify speaker intent as they encode semantic and discursive meaning: accent is a medium, through which we project individual style and signal our relationship to interlocutors. Even more broadly, it reflects social identity along various categorical lines. (Moyer, 2013: 19)

One of the main challenges for accent research is why it is that some learners are capable of reaching a native level in a L2 despite having a late start (post childhood). Moyer (2013: 81) argues that exceptional learners 'all felt a deep connection to the target language and all took a conscious reflective approach to the learning process, regardless of their different amounts of formal instruction … all directed their attention to the sounds of the language in order to emulate native speakers'. What is crucial in such attainment is social interaction with natives, but clearly that is not sufficient, since many have that experience but do not achieve a nativelike accent, presumably because they choose not to.

In recent years, foreign accents have required forensic analysis in legal settings because they are regarded as establishing national and ethnic identity (Eades, 2005). The notion that instruction in phonology can improve accent has also been promulgated, making accent reduction a focus of

classroom instruction (Harding, this volume; Isaacs, 2014). Moyer challenges this notion with two claims:

(1) *If accent in a second language is such an individual skill and much of what contributes to it lies beyond conscious practice, how consequential can explicit training be?*
(2) *Can the classroom address phonological fluency in a comprehensive way, providing enough contextualized practice to promote real authenticity?* (Moyer, 2013: 147)

Moyer's honest conclusion is that she is doubtful whether instruction can/does have such an effect. It is all very well, she points out, aiming at intelligibility if we are not quite sure what intelligibility is. Moyer (2013: 172) concludes, 'Overall language proficiency should be disambiguated from accent', but ruefully admits that accent discrimination is on the rise in the US and the UK, echoing Honey (1989). Globalization and migration bring in their own revenges. And this is where Moyer's honesty is at its most affecting: there is no alternative to input and practice. 'Few among us are willing to go to … extremes, leaving our linguistic (and emotional) comfort zone for the sake of fully adopting the sounds of a different culture' (Moyer, 2013: 178). But most honest of all is Moyer's (2013: 178) last paragraph, where she evokes Gardner's (1979) notion that becoming nativelike in accent requires 'desire plus effort sustained over the long term'. Here, indeed, is the serious challenge for pedagogy: how to foster desire.

If the spoken language is to be tested, there has to be an agreed upon model. In most situations, this will be the local prestige 'class' accent, and for post-colonial situations, it is likely to be the colonial prestige. This was where I somewhat ruefully arrived after my West African project described above – a position since confirmed by evidence from India, Nepal and Singapore (Davies, 2013). My conclusion to the accent–native speaker assessment relation is that if the spoken language is to be assessed and if accent is one of the variables under test, then the native speaker in its idealized representation as a prestige variety is needed as model and goal.

Note

(1) Alan Davies wrote this candid commentary on the native speaker in March 2015, shortly before his death in September 2015 and prior to the publication of this edited volume. His original words, likely representing his latest thinking and scholarly output on the topic, appear here in their original form with minimal editing, with publication in this edited volume supported by his family. His voice and thinking and ideas, crystalized in his many works and reflected in the discursive, nearly conversational style of this chapter, will continue to resonate for a long time to come.

References

Bloomfield, L. (1927) Literate and illiterate speech. *American Speech* 2, 432–9. Reprinted in D. Hymes (ed.) (1964) *Language in Culture and Society* (pp. 391–6). New York: Harper & Row.

Braine, G. (1999) From the periphery to the center: One teacher's journey. In G. Braine (ed.) *Non-native Educators in English Language Teaching* (pp. 15–27). Mahwah, NJ: Lawrence Erlbaum.

Canagarajah, S. (1999) *Resisting Linguistic Imperialism*. Oxford: Oxford University Press.

Chomsky, N. (1985) *Knowledge of Language: Its Nature, Origin and Use*. New York: Praeger.

Cook, V. (1999) Going beyond the native speaker in language teaching. *TESOL Quarterly* 33, 185–209.

Crystal, D. (1997) *Dictionary of Linguistics and Phonetics*. Oxford: Blackwell.

Davies, A. (2005) *A Glossary of Applied Linguistics*. Edinburgh/Mahwah, NJ: Edinburgh University Press and Lawrence Erlbaum.

Davies, A. (2013) *Native Speakers and Native Users: Loss and Gain* (2nd edn). Cambridge: Cambridge University Press.

de Saussure, F. (1916) Nature of the linguistic sign. In C. Bally and A. Sechehaye (eds) *Course in General Linguistics*. New York: McGraw-Hill.

Eades, D. (2005) Applied linguistics and language analysis in asylum seeker cases. *Applied Linguistics* 26 (4), 503–526.

Elder, C. and Davies, A. (2006) Assessing English as a lingua franca. *Annual Review of Applied Linguistics* 26, 282–301.

Ferguson, C. (1983) Language planning and language change. In H. Cobarrubias and J. Fishman (eds) *Progress in Language Planning: International Perspectives* (pp. 29–40). Berlin: Mouton.

Gardner, R. (1979) Social psychological aspects of second language acquisition. In H. Giles and R. Sinclair (eds) *Language and Social Psychology* (pp. 193–220). Baltimore, MD: University Park Press.

Halliday, M.A.K. (2003) Written language, standard language, global language. *World Englishes* 22 (4), 405–418.

Holliday, A. (2005) *The Struggle to Teach English as an International Language*. Oxford: Oxford University Press.

Honey, J. (1989) *Does Accent Matter? The Pygmalion Factor.* London: Faber & Faber.

Hyltenstam, K. and Abrahamsson, N. (2000) Who can become native-like in a second language? All, some or none? On the maturational constraints controversy in second language acquisition. *Studia Linguistica* 54, 150–166.

Hyltenstam, K. and Abrahamsson, N. (2003) Maturational constraints in SLA. In C.J. Doughty and M.H. Long (eds) *The Handbook of Second Language Acquisition* (pp. 539–588). Malden, MA: Blackwell.

Isaacs, T. (2014) Assessing pronunciation. In A.J. Kunnan (ed.) *The Companion to Language Assessment* (pp. 140–155). Hoboken, NJ: Wiley-Blackwell.

Jenkins, J. (2005) Implementing an international approach to English pronunciation: The role of teacher attitudes and identity. *TESOL Quarterly* 39 (3), 535–543.

Kachru, B. (1992) *The Other Tongue* (2nd edn). Urbana, IL: University of Illinois Press.

La Rochefoucauld, F., duc de. (1822) *Réflexions, sentences, et maximes morales de la Rochefoucauld*. Paris: Chez Lefebvre.

Medgyes, P. (1992) Native or non-native: Who's worth more? *ELT Journal* 46 (4), 340–349.

Moyer, A. (2013) *Foreign Accent: The Phenomenon of Non-native Speech*. Cambridge: Cambridge University Press.

Mugglestone, L. (1995) *Talking Proper: The Rise of Accent as Social Symbol.* Oxford: Clarendon Press.

Seidlhofer, B. (2004) Research perspectives on teaching English as a lingua franca. *Annual Review of Applied Linguistics* 24, 209–239.

Wolfram, W. (2004) Social varieties of American English. In E. Finegan (ed.) *Language in the USA: Themes for the Twenty-first Century* (pp. 58–75). Cambridge: Cambridge University Press.

11 Variation or 'Error'? Perception of Pronunciation Variation and Implications for Assessment

Stephanie Lindemann

Introduction

All thriving languages include variation: variation in lexical choices, in grammar and in pronunciation. While research in second language acquisition tends to focus on learners' interlanguage, or their variation from a perceived native speaker norm, research in sociolinguistics traditionally investigates patterns in language variation among native speakers. At the most basic level, language may vary in association with social factors relevant to individuals such as social class and region of origin, or with factors relevant to the social situation of language use, such as level of formality or desire to accommodate to one's interlocutor.

Approaches to pronunciation assessment, whether for in-class formative assessment or for high-stakes exams, may make reference to how well pronunciation matches a particular norm – as in accuracy-based assessment – or by whether or how easily a given listener understands the speaker, as in intelligibility- or comprehensibility-based assessment. This chapter discusses aspects of variation and related listener perception that complicate both types of assessment, especially when assessment is couched within a deficit model of the nonnative speaker. Such a model may be expressed through a focus on 'errors' or on purely speaker-focused assessments of unintelligibility.

This chapter is organized as follows. First, I discuss variation in English pronunciation, focusing on variation used by native speakers whose language may be widely perceived as standard. While the term 'standard' implies uniformity, it is typically applied to language that is not necessarily uniform, but

rather is spoken by comparatively powerful social groups whose language is widely regarded as 'correct' or as a good model for language learners. However, listeners are frequently unaware of variation in what they consider standard English. Then, I compare these observations about perception of 'standard' native speech with perception of variation in nonnative speech. The existence of variation in both native and nonnative speech, paired with the role of expectations about the speaker in the perception of that variation, call into question the seemingly straightforward assessments of a speaker's pronunciation as 'standard' or not. Next, I consider the role of attitudes in speech perception, showing that listeners' attitudes towards speakers, as well as their expectations about how they will sound, have also been shown to affect their comprehension of the speaker and their assessments of the speaker's intelligibility. I discuss implications for assessment in the final section.

Variation and Perception of Variation in Native English Pronunciation

Most sociolinguistic research that investigates variation in pronunciation across native English dialects – variation that indexes, or points to, speakers' social class, ethnicity, gender, age or region of origin – focuses on segmental differences, such as variation in vowel systems. Research on variation in suprasegmentals such as intonation and word stress is far less common. Thus, this review focuses on the significance of variation in segmentals. However, it should be noted that even within a given country there is variation in the pronunciation of suprasegmental features. For example, speakers of different regional varieties in the US may stress either the first or the second syllable in words like *July*, *hotel* and *theatre* (Wolfram & Schilling-Estes, 2006). Intonation patterns may also vary based on race (see Thomas & Reaser, 2004, for a review of research comparing European-American and African-American intonation patterns) and other factors. For example, an analysis of uptalk, the use of rising 'questioning' intonation for statements, in game show data found that patterns of use varied depending on gender, race, interlocutors, the status of the statement (whether it was an initial response or correcting another's response), and the contestant's level of success in the competition, with these factors interacting (Linneman, 2013). Thus, it is possible that pronunciation assessment that pays relatively more attention to suprasegmentals may be complicated by some of the same challenges we will see for individual phonemes (as suggested by Sewell for intonation of questions, this volume), but there is as yet relatively little data from which to draw implications.

Turning to pronunciation of segmentals, it is important to recognize the substantial variation within native varieties, even within what may be perceived as a standard variety such as 'General American' and described by

people from the US as 'average', 'normal' or 'correct' (Preston, 2008). This section will focus primarily on examples of such unstigmatized and largely non-salient variation to demonstrate that a pronunciation model such as 'General American' is better thought of as a sometimes convenient but generally misleading fiction rather than as sociolinguistic reality.

The US Midwest in general, and Ohio in particular, are frequently cited by laypeople as examples of where so-called General American is spoken. It is first worth noting that sociolinguistic descriptions of Midwest English (Frazer, 2006; Gordon, 2006) and Ohio English (Flanigan, 2006) all point out that the language spoken in these areas is by no means uniform. Perhaps more surprisingly, many speakers from the US Midwest who would be identified as General American speakers – including highly educated, middle-class speakers – do not make all of the vowel contrasts taught as part of 'General American', and may even pronounce most vowels differently from what is usually presented as a General American vowel system.

In terms of vowel contrasts, many speakers who might be identified as standard speakers do not distinguish the vowels used in the words LOT and THOUGHT, so that they pronounce *cot* and *caught* the same way (i.e. low-back merger). Such speakers are usually surprised to learn that other English speakers in the US do make this distinction and that their own variety thus may be seen as 'lacking' a vowel, in spite of its status as a 'standard' variety. Although miscommunication does occur between speakers who distinguish the vowels and those who do not (Labov, 2010), the lack of a distinction is not stigmatized, or even noticed in the absence of miscommunication.

Also common in these Midwestern varieties that may be identified as 'General American' is a series of changes in the vowel system that can result in the intended vowels sounding like completely different phonemes to speakers of other varieties. These changes include pronunciations of *bet* like *but* (i.e. /ɛ/ as closer to /ə/), *but* like *bought* (/ə/ more like /ɔ/), *bought* like *bot* (/ɔ/ more like /ɑ/), *bot* like *bat* (/ɑ/ as closer to /æ/), and so on, a phenomenon called the Northern Cities Shift (NCS), since it is mainly speakers in urban areas of the north that participate in these changes. What is particularly interesting is that these rather dramatic vowel changes go largely unnoticed, in spite of the fact that they have been documented as being repeatedly misunderstood by speakers of other varieties of US English (Labov, 2010). These differences are also not associated with much social evaluation, with no differences found between listeners' evaluations of the intelligence, education or trustworthiness of a speaker with NCS vowels compared to one with unshifted vowels (Labov, 2010). As Gordon (2006: 110) points out, 'As long as Midwesterners are viewed as average, boring or otherwise nondescript, their speech will be seen through the same prism'.

Midwesterners themselves may fail to hear this vowel shift produced by supposed General American speakers even when they are specifically focusing on vowel pronunciation. Niedzielski (1999) presented listeners from

Detroit, Michigan (in the Midwest) with sentences produced by a speaker who was also from Detroit and used NCS vowels. The listeners were asked to match the vowel of a keyword in the sentence to one of three computer-synthesized vowels. When they were told that the speaker was from Detroit, they seldom chose the shifted vowel most similar to what the speaker actually produced, instead choosing a more 'standard' or even a 'hyperstandard' vowel – one that was shifted in the opposite direction of the actual shift. Additionally, the speaker they heard also used Canadian-raised vowels, such that the vowel in *house* started with a schwa-like vowel rather than with /a/. (This feature is often parodied, not very precisely, by quoting a Canadian speaker as saying something like *aboot* for *about*.) Speakers from Detroit also commonly use Canadian raising, but it is stereotyped as being used only by Canadians. In fact, the listeners were significantly more likely to identify this vowel accurately when they were told that the speaker was Canadian than when they were told she was from Detroit. Thus, their poor matching skills did not result from an absolute inability to hear the differences between the vowels, but rather from their own preconceived notions of what the speaker would sound like.

In sum, non-stigmatized varieties of US English may (1) lack a vowel distinction found in other varieties of US (and other) English, and (2) involve a number of differences in the vowel system from what may be expected from 'General American' English. Certainly, when writers refer to 'General American', they refer to vowels that have not been affected by the NCS, even though speakers in areas most frequently identified as the source of 'General American' are increasingly pronouncing their vowels according to the NCS. When North American English is presented as a pronunciation model (e.g. in Celce-Murcia *et al.*, 2010), Midwesterners may assume that this matches their own pronunciation, even if they actually pronounce NCS vowels. Of course, vowel systems vary still more when speakers from different countries are considered, with different pronunciations as prestigious or stigmatized in each.

Variation in consonant pronunciation in native varieties tends to be more subtle, with fewer examples of consonants being perceived as different phonemes by speakers of different dialects of US English, although this does occur. This discussion will focus on consonant variation within native English that is sometimes treated as problematic when produced by nonnative speakers, including pronunciation or omission of /r/, as well as varying pronunciations of /l/, /t/, and interdental fricatives (the 'th' sounds).

Of the consonants, only variation in pronunciations of interdental fricatives may be widely stigmatized; these may be pronounced as stops (/t/, /d/) or as labiodentals (/f/, /v/). I am not aware of any data suggesting that these pronunciations are regularly associated with miscommunication among native speakers in inner circle countries, although Deterding (2005) has noted that pronunciation of interdentals as labiodentals at the beginning of

a word sometimes caused difficulty for highly proficient, possibly native, English-speaking listeners from Singapore.

Other variation, while not universally stigmatized, may be stigmatized under certain conditions. For example, evaluation of /r/-pronunciation after vowels differs depending on the speaker and their location, with /r/ absence seen as prestigious in the UK but stigmatized in most US dialects (Wolfram & Schilling-Estes, 2006). The pronunciation of /t/ as a glottal stop may go completely unnoticed depending on its location in the word as well as on the speaker's location in the world. For example, most students in my classes in the US have been surprised to learn that they or others did not pronounce a [t] in the word *important* at all, pronouncing both /t/ sounds as glottal stops. On the other hand, they easily hear the glottal stop when it is pronounced in a word like *butter.* They may identify this as a feature of a British accent, in which case they are likely to identify it as prestigious, or more specifically as a Cockney accent, in which case they may identify it as stigmatized. Meanwhile, in the UK, the use of a glottal stop in a word like *butter* has traditionally been stigmatized, but is becoming increasingly accepted (Przedlacka, 2002).

A final example of consonant variation among native English varieties that may be addressed in pronunciation courses for English learners is the pronunciation of /l/ as velarized – with the back of the tongue raised in addition to the tip touching the alveolar ridge behind the teeth, also called 'dark' /l/ – or even as a vowel. While British English speakers typically pronounce /l/ as unvelarized, or 'light', before a vowel, as in *leaf*, and velarized only after the vowel in a given syllable, as in *feel*, US speakers are less likely to make this distinction, producing all /l/s as more or less velarized (Ladefoged & Johnson, 2011). Meanwhile, for speakers in many areas of the US, /l/ at the end of a syllable may be produced without the tongue tip touching the alveolar ridge at all, such that syllable-final /l/ is becoming a vowel (Labov, 2010; Ladefoged & Johnson, 2011). None of the variation in pronunciation of /l/ appears to be particularly salient to native speakers in the US. For example, /l/ pronunciation as a vowel is not commonly the subject of overt commentary in spite of the fact that, like variability in /r/ pronunciation, it has been documented as a factor in misunderstandings among native speakers of different varieties of US English (Labov, 2010), with differences in its pronunciation leading to perception of a different word from that intended by the speaker, such as *balance* perceived as *bounce*.

Another source of variation has more to do with natural pronunciation of connected speech, regardless of variety. Various types of assimilation and reduction present in native speech are well known to pronunciation teachers, who may teach such forms and assess their students' ability to produce or understand them (Sewell, this volume). However, some forms of lenition, in which segments are omitted or produced with less closure than they are typically produced when the word is pronounced in citation form, may be less familiar. These changes may be quite striking, resulting in sounds not

considered part of the English phonological system, such as velar fricatives or approximants (Shockey, 2003; Simpson, 2013). For example, a Midwestern US speaker I recorded pronounced *tiger* without closure on the /g/, so that the word actually sounded more like *tire*, a pronunciation that I only noticed when I heard the recorded word out of context.

Perception of 'Nonnative' English Variation

As we have seen, there is variation within native English varieties, especially in vowels, but also in consonants to some extent. Much of this variation goes unnoticed when it is produced by a speaker who is perceived as a 'standard' speaker – in the US, this would be especially true for White, middle-class speakers from the Midwest. In fact, the Niedzielski (1999) study discussed above has already demonstrated that the same pronunciation features may be perceived differently depending on who is believed to be using them and what stereotypes exist about the perceived speaker. In that case, perception of vowels undergoing Canadian raising were perceived accurately if the speaker was believed to be Canadian, but inaccurately if the speaker was believed to be from the US. Similarly designed studies have found parallel results with differing perception of vowels based on apparent nationality (this time in New Zealand; Hay *et al.*, 2006a), or perceived age (Hay *et al.*, 2006b). Relatively subtle differences in consonant perception, specifically in /s/ and /ʃ/ sounds, as in *sod* and *shod*, have also been found to be affected by the perceived gender of the speaker (Strand, 1999).

The preceding studies all involved perception of native English varieties in which none of the various pronunciations were perceived as 'errors' per se. However, when listeners were presented with words containing NCS vowels (such as *block* pronounced like non-NCS *black*) out of their original spoken context, even NCS speakers usually misidentified the word *block* as *black* (Labov, 2010). Even when the word was presented with the rest of the utterance 'senior citizens living on one _', a third of non-NCS speakers and a quarter of NCS speakers still did not recognize the word as *block* (Labov, 2010). Thus, this pronunciation would likely be regarded as an 'error' if such a judgement were based on the actual pronunciation and its interpretation by native speakers alone, but, as mentioned above, outside artificial experimental conditions NCS speakers do not even notice their 'error', and neither do speakers of other varieties in the US.

When such variation is produced by English learners or users, or even by native speakers from places like India or Singapore, however, it is much more likely to be interpreted as a pronunciation 'error'. To take an anecdotal example that parallels the NCS vowel perception described above, Tom Horne, then the superintendent of Arizona schools, attempted to explain the state's policy on English teacher proficiency by arguing in a news interview

that '… if you- if you mispronounce words to the extent that they sound like other words, um … you shouldn't be teaching kids English' (Kidd, 2010). In the context of that interview and others, the superintendent was explicitly referring to nonnative pronunciations, giving a native Spanish-speaking teacher's pronunciation of *comma* as *coma* as an example.

Of course, such an example is likely to be familiar to teachers and researchers of pronunciation, and learners of English might well improve their intelligibility for a range of listeners by reliably distinguishing between the vowels in *comma* and *coma*. The point is that the simple identification of one word as 'sounding like another word' should not necessarily be a problem, or be regarded as such. Jenkins (2000) has argued that, in communication among users of English, using a specific vowel quality is less important to comprehensibility than having consistent vowel quality. Her argument is supported by researchers demonstrating that listeners do quickly adapt to different varieties – as would be necessary for the NCS to go largely unnoticed. For example, Clarke and Garrett (2004) found that listeners' processing time for a nonnative speaker's accent decreased significantly with one minute of exposure to that speaker. The fact that the LOT-THOUGHT merger also goes unnoticed in the US suggests even that some mergers are not entirely detrimental to overall successful communication, although miscommunications may arise from differences among speakers' vowel systems, as noted above. Here the point is that the perceived seriousness of the 'error' – and indeed, whether it is perceived as an error at all – may have as much to do with who the speaker is perceived to be, as with the actual pronunciation or even with miscommunications that may occur.

There is also experimental evidence for differences in perception depending on whether speakers are believed to be native or nonnative. Hu and Lindemann (2009) used methodology similar to Niedzielski (1999) to investigate Cantonese speakers' perception of word-final stops such as the /k/ in *book*. They found that Cantonese English speaking undergraduates describe 'incomplete' or 'deleted sounds', especially at the ends of words, as a way in which their own speech contrasts with American English, implying that word-final stops in American English are not 'incomplete' (or unreleased). When these speakers perceived English keywords with word-final stops in the context of a sentence, they were significantly more likely to match the keyword to a word with an 'incomplete' (unreleased) stop if the speaker was identified as Cantonese, but as a fully pronounced, even aspirated stop if the speaker was identified as being from the US. In both cases the actual speaker was from the US, and she often did not release word-final stops in sentence context. In other words, the listeners' perception was more accurate when they thought the speaker was Cantonese, but they were likely to identify that perception as a quality of less than 'perfect' English.

The difference between released and unreleased stops is a relatively subtle one, albeit one that is remarked on and even stereotyped as a feature of

Cantonese English. We might expect perception of other consonant distinctions to be more straightforward. For example, /p/ and /f/ sounds are heard as different phonemes and are not usually expected to be confused by native English speakers. In addition, the distinction between them carries a high functional load, differentiating many words (Brown, 1991; Munro & Derwing, 2006). However, Labov (2010) notes that miscommunications that were observed between speakers who merge the LOT and THOUGHT vowels and those who do not merge them included numerous examples of mishearing the word *coffee* as *copy* and vice versa, even when the context lent itself to the correct interpretation. He suggests that the non-salience of the distinction between /f/ and /p/ when they appear between vowels allowed the perception of the vowel to dominate in the identification of the specific word. This suggests that even more significant consonant distinctions are not always perceived very distinctly by native speakers, thus opening up the possibility that they may be heard differently depending on whether the speaker is identified as native or nonnative. In the former case, consonant differences may be less likely to be noticed; in the latter, consonant differences may be more likely to be noticed – or even imagined when not present.

Evidence of such differential noticing of native compared to nonnative pronunciation comes from data I have collected from 59 undergraduate native speakers of US English who were asked to listen to an English language text spoken by native speakers of US English, Italian, Korean and Mexican Spanish, and to adjust a written version of the text to reflect how each speaker sounded. The speakers were identified to participants only by speaker number rather than by nationality. Analysis of which words participants respelled showed that they were more likely to respell words produced by nonnative English speakers, even when those speakers pronounced the words identically to the native speaker. For example, unstressed vowels are frequently pronounced as schwa, so that the indefinite article *a* might be respelled as *uh*. While such vowels were respelled a small percentage of the time, the percentage of each nonnative speaker's respelled reduced vowels was at least twice that of the native speaker's. (Of all possible reduced vowels produced by each speaker, listeners respelled 4% of the Korean English speaker's vowels, 3% of the Mexican's, 2% of the Italian's and 0.75% of the American's). Perhaps more significantly, respellings of words like *together* as *togeder* were sometimes used for nonnative speakers when they produced the interdental fricative accurately; this never occurred in respellings of native speech.

The respelling task also demonstrated that it was frequently difficult for listeners to pinpoint even fairly straightforward pronunciation differences, suggesting that listeners may have a global sense of nonnative pronunciation even in a task that focuses on detail. The listeners were able to control playback of the sound file so that they could listen in short chunks and hear each part as often as they needed. Nevertheless, in many cases in which they

apparently perceived nonnative features produced by speakers, they perceived them inaccurately. For example, a Korean speaker pronounced *methods* similar to *message* near the beginning of the sound file. Because his pronunciation resulted in a different, easily recognizable English word, we might expect it to be relatively easy for participants to respell in a way that reflected his pronunciation. However, respellings included *mezods, meshods* and *mehods*, with only about 40% of participants transcribing it in a way that reflected the pronunciation of both the *th* and the *ds*. Similarly, the same speaker used an r-less pronunciation of *world* at the end of the first sentence, after which he paused. In spite of this rather straightforward pronunciation in a spot that was comparatively easy to hear, respellings included *word, worll* or even *wurld*, this last example perhaps in recognition of some 'non-standard' feature, but actually respelling the word in a way that would not likely reflect a different pronunciation from its standard spelling. Again fewer than 40% of participants identified the r-less pronunciation without also making inaccurate changes such as omitting the /d/. Thus, in the absence of phonetic training, when a task asks listeners to focus on specifics and makes it relatively easy to do so, they may still have difficulty with perception of those specifics.

Taken together, these findings suggest that identifying specific pronunciations, not to mention identifying 'errors', is not as straightforward as it may initially seem. The same sounds may be heard differently depending on the listeners' skill and even on who they perceive the speaker to be. The respelling data, as well as the results from Hu and Lindemann's (2009) study, suggest that the language of nonnative speakers may undergo greater scrutiny than that of non-stigmatized native speakers, making identification of native pronunciations (including reduced vowels and unreleased stops) as 'errors' more likely.

Bias Against Nonnative Speakers

Over-perception of nonnative speech features identified as 'errors' may be exacerbated by issues of systematic bias against (perceived) nonnative speech, especially against that spoken by non-White speakers. We see some evidence in the respelling data above in which a Korean speaker's reduced vowels were respelled twice as often as an Italian's, while the Italian's reduced vowels were respelled more than twice as often as a native US English speaker's vowels, whose reduced vowels were largely treated as 'normal' and thus not in need of respelling, or not noticed at all. This order of 'nonnativeness' is also matched by data on listeners' ratings of how correct, pleasant and friendly each speaker sounded: the Korean speaker was rated most negatively on all traits, while the US English speaker was rated most positively on correctness and was rated similarly to the Italian English speaker on pleasantness and friendliness.

Bias against nonnative speakers is well-documented in numerous verbal guise studies, which present multiple speakers all reading the same text and ask listeners to rate the speakers on status qualities like intelligence and education, as well as on social attractiveness qualities like friendliness or kindness. While results on social attractiveness qualities vary, nonnative speakers are nearly always evaluated lower on status traits than are native speakers, both by native listeners (Ball, 1983; Eisenchlas & Tsurutani, 2011; Lindemann, 2003; Nejjari *et al.*, 2012; Ryan & Bulik, 1982; Ryan *et al.*, 1977) and by nonnative listeners (Chiba *et al.*, 1995; Dalton-Puffer *et al.*, 1997; He & Zhang, 2010; McKenzie, 2008, 2010; Xu *et al.*, 2010; Yook & Lindemann, 2013).

In even clearer cases of bias, listeners have rated the same speakers differently depending on how the speakers were explicitly identified. For example, Buckingham (2014) found that Omani listeners rated a Pakistani English speaker much more highly when he was identified as being from the UK; their ratings of other nonnative speakers also tended to be higher when those speakers were identified as native. For cases in which listeners already recognize a speaker as nonnative, explicit identification of her as being of their own nationality might result in more positive ratings. For example, Yook and Lindemann (2013) found that Korean listeners rated a Korean English speaker significantly more highly on status traits when she was explicitly identified as Korean than when they were asked to guess where she was from, although she was still rated significantly more negatively than all explicitly identified native speakers.

These differential ratings of speakers depending on how they are identified, as well as the tendency towards negative evaluations of nonnative speech overall, are particularly important to the current discussion because negative attitudes towards speakers have been found to be associated with erroneous perception of nonnative accent where none is present (Kang & Rubin, 2009). Rubin (1992) found that, when US undergraduates listened to a recording of a lecture produced by a native English speaker, they scored lower on a comprehension test and rated the speaker's accent as less nativelike if the speaker was visually identified (via a photograph of the supposed speaker) as Asian rather than if she was identified as Caucasian. Kang and Rubin (2009) found that such an effect was more likely to occur among listeners who could be identified as having negative attitudes towards nonnative speech.

Most of these studies were conducted with undergraduates or others who have not chosen to dedicate their careers to working with nonnative speakers. We might expect language teachers to have more positive views of nonnative speakers, leading them to be less susceptible to biased perception. In addition, there is some evidence that students who are taking a linguistics class may rate nonnative accents more positively than do the general undergraduate population. Eisenchlas and Tsurutani (2011) found that students who were taking a linguistics class and studying at least one foreign language rated an Argentinian Spanish speaker as highly in competence and integrity

traits as they rated a native speaker from Australia, although they still rated native speakers of Japanese, Korean, Farsi (and to a lesser extent Italian) more negatively than the Australian on competence traits. Even more strikingly, Pantos and Perkins (2013) found that their US undergraduate study participants, about 80% of whom were native English speakers, showed explicit preference for a native speaker of Korean over a native speaker of US English in terms of which 'witness' they would side with in a malpractice case, with measures of speaker traits showing no significant differences. Nearly half of these participants were registered in a sociolinguistics class that addressed issues of language discrimination. Thus, interest in linguistics and different languages as well as education about sociolinguistic issues may be associated with less bias against nonnative speakers. We might expect language teachers to have more in common with these groups than with the general population. In fact, Litzenberg (2013) found that pre-service ESL/EFL teachers rated native speakers and advanced nonnative speakers equally on many traits including competence, although intermediate-level nonnative speakers were usually rated lower. These pre-service teachers did not rate native and nonnative speakers differently on education at all, although this may be because all speakers were explicitly, if somewhat subtly, identified as students.

These findings, while encouraging, hardly guarantee that language teachers are completely immune to the bias against nonnative speech in the wider society. In particular, Pantos and Perkins (2013) found that, while their participants showed explicit preference for a Korean speaker over a native English speaker, they still showed an implicit bias in favour of the native speaker. This implicit bias was detected by the Implicit Association test (Greenwald *et al.*, 1998), in which participants' reaction time is measured for how quickly they associate Korean (for example) with positive traits in comparison to how quickly they do so for negative traits. Thus, we might expect good intentions to allow language teachers to make more positive choices with respect to nonnative speech when there is opportunity for reflection. However, automatic, implicit processes such as those involved in speech perception may still be subject to society-wide biases.

The degree to which implicit biases may affect the perceptions of those who are explicitly appreciative of nonnative speakers remains an open question. One area in which teacher training may be considered to have a mixed effect is reflected in the finding that pre-service teachers near the end of their training were more critical of nonnative language features found in examples of successful communication than those near the beginning of their training. Specifically, Litzenberg (2013) found that students in BA TESOL programmes were more lenient in their assessments of nonnative speakers' language ability and level than were students in MA programmes, and new MA students were more lenient than those near the end of their MA programme on assessments of level. While this finding could be positive in the sense that TESOL training might assist listeners in pinpointing the specific difficulties a

language learner might be having, it also suggests that a focus on close analysis of language can lead listeners to become more aware of language features that may not interfere with speakers' ability to communicate, including some of the otherwise non-salient variation in specific sounds. (Indeed, Kennedy *et al.*, current volume, found that two raters who had more study in phonetics and phonology showed greater orientation to a native speaker model than did two with less such study.)

Implications for Assessment

We have seen that there is substantial variation in native varieties of English, even within what is considered 'standard', but that this variation goes largely unnoticed, even though it sometimes leads to miscommunication among speakers of different varieties. Such variation becomes noticeable and sometimes perceived as 'erroneous' when it is produced by nonnative speakers; in some cases, 'errors' or other types of variation may even be imagined or otherwise misperceived when the speaker is believed to be a nonnative speaker. Such misperceptions may be linked to societal biases, but over-perception of variation could also stem from linguistic training that focuses closely on nonnative speech production, especially if such training does not include closer investigation of the variation naturally present in non-stigmatized native varieties. Teacher training may assume that preservice teachers already have an understanding of what native speakers do and focus on nonnative speakers' 'mistakes' without acknowledging that native speakers may also produce velar fricatives, 'incomplete' word-final stops, and words that 'sound like other words'.

These sociolinguistic findings suggest that an accuracy-based measure of pronunciation is likely to be somewhat arbitrary, especially for pronunciation of vowels, which varies widely among native speakers. Instead, a focus on the speaker's consistency in the vowel qualities used, regardless of whether these qualities match a particular native variety, would be more relevant, as Jenkins (2000) has suggested. In the case of both vowels and consonants, 'errors' may be over-perceived when a speaker is believed to be nonnative. Thus, it may not be meaningful or reliable to talk about accuracy or accentedness as a measure of pronunciation. Instead, as many have argued, there is a need for emphasis on intelligibility, which does not necessarily correlate with perceptions of nonnative accent (Munro & Derwing, 1995).

Of course, assessing intelligibility comes with its own challenges, and does not eliminate the possible effects of bias discussed above. The undergraduates in Rubin's (1992) and Kang and Rubin's (2009) studies, who had difficulty in understanding a speaker when they erroneously believed that speaker to be nonnative, are a case in point. As Rajagopalan (2010) has argued, intelligibility, although often treated as a neutral term, is one that is

meaningless without an implied evaluator, one to whom the speech is intelligible or otherwise. Thus, the role of the assessor cannot be ignored in the assessment of intelligibility. At the very least, evaluators' familiarity with variety plays a role in their assessments of pronunciation (Ballard & Winke, this volume; Browne & Fulcher, this volume; Carey *et al.*, 2011).

For high-stakes testing in particular, it is thus relevant for rating scale descriptors referring to intelligibility to at least specify whether the intended listener is familiar with the accent, e.g. with higher scores for speakers who are intelligible even to listeners who are unfamiliar with the accent (Breiner-Sanders *et al.*, 2000). Making such an assessment presents its own challenges, however, since it implies a need to find trained examiners who are familiar with any tested accent as well as others who are unfamiliar; otherwise observers must feign a scenario in which they have a different level of familiarity and extrapolate the speaker's intelligibility for the imagined listener.

One approach to addressing this difficulty, although appearing to return to an accuracy-based measure and sometimes using deficit-oriented terminology, may provide a somewhat more systematic way to assess intelligibility by attempting to specify what kinds of variation are not problematic for successful communication (Jenkins, 2000) or comprehensibility (Munro & Derwing, 2006). Ideally, this determination is made based on what variation is found in successful communication with a wide range of interlocutors who demonstrate goodwill to the speaker (Lippi-Green, 2012) and motivation to communicate. In Jenkins's data, pronunciation of interdental fricatives (voiced and voiceless 'th' sounds) as stops (/d/ and /t/) or as alveolar fricatives (/s/ and /z/) was not associated with any communicative difficulty in a wide range of interactions between English users; nor was most variation in vowel quality associated with communication breakdowns as long as vowel quality was consistent for a given speaker. If this pattern is pervasive across interactions, this would suggest that such variation could be accepted even at the highest levels of speaking proficiency. There would then be less relevance attached to whether assessors accurately perceived this variation. Other possible factors in the relationship between the pronunciation of specific sounds and intelligibility include where the sound is in the word (Bent *et al.*, 2007) and the importance of the word in understanding the utterance (Isaacs, 2008).

Of course, there were some individual sounds in Jenkins's (2000) study that were deemed crucial for successful communication and that have been triangulated by subsequent work (e.g. Munro & Derwing, 2006) as being important for comprehensibility. If assessment of intelligibility were to focus on the production of those specific sounds, attention would still need to be paid to whether they are perceived accurately by the person assessing pronunciation. An encouraging, although indirectly related finding from Kang and Moran (2014) is that speech samples that had been rated at higher levels of speaking proficiency in the Cambridge ESOL General English examinations included alternate pronunciations of interdental fricatives and several

vowels; one of the alternate vowel pronunciations is also found in the NCS (specifically, a change in the vowel used in DRESS from [ɛ] to [ə], so that, for example, *restaurant* is pronounced as *rustaurant*). In other words, pronunciations that deviated from a 'General American' norm, but that we would not expect to affect intelligibility, did not seem to prevent speakers from getting a higher proficiency score.

Thus, examiners' assessments of the speaker's intelligibility and some measure of their own familiarity with the accent could perhaps be supplemented with a secondary analysis of specific sounds that have been found to be especially relevant to successful communication. Such an approach would require further research regarding a complete picture of which sounds are most crucial (if indeed this can be determined across a wide range of speaker and listener backgrounds) in order to be used for high-stakes testing.

In addition to assessors' familiarity with speakers' varieties, their beliefs about the speaker and even unconscious biases against certain groups may also be relevant to testing outcomes. This suggests a need to investigate any attitudinal biases that may influence examiners' ratings. If such an influence is found, a screening instrument would need to be developed that could be given to examiners for high-stakes tests, possibly utilizing a type of Implicit Association test (briefly described in the section above) for a variety of speaker backgrounds.

Conclusion

Sociolinguistic findings suggest that using a deficit-based approach to assess pronunciation in terms of 'errors' is problematic for reasons of both ideology and precision. Defining a particular pronunciation as an error implies a 'correct' version, promoting a standard language ideology (Lippi-Green, 2012) that privileges those whose language is viewed as standard in spite of variation in that language, to the detriment of those who are not viewed as standard speakers, in some cases based purely on their appearance. Because the standard is always implied rather than an objective reality (similar to ideas of 'natural' speech; see Harding and Sewell, both in this volume), defining speech as an 'error' because it is perceived as deviating from this vaguely defined standard also lacks precision. This is not unlike referring to a speaker as 'having an accent' without specifying what kind, since all speakers have an accent.

This argument can also be applied to measures of intelligibility, which are often treated as measures purely of the speaker, while ignoring the listener's role (but see Zielinski, 2008, for an excellent counter-example). When the listeners to whom a speaker is or is not intelligible are specified, the term gains more precision, as well as presenting a less deficit based view of the speaker. When the relevant groups of listeners are identified, those listeners'

possible biases may be taken into account. Most immediately, we can take possible biases into account by not assuming assessments of intelligibility to be neutral measures of the speaker alone. In the longer term, the role of such biases in intelligibility requires more investigation and, ultimately, ways of reducing such biases must be developed.

References

Ball, P. (1983) Stereotypes of Anglo-Saxon and non-Anglo-Saxon accents: Some exploratory Australian studies with the matched guise technique. *Language Sciences* 5 (2), 163–183.

Bent, T., Bradlow, A.R. and Smith, B.L. (2007) Segmental errors in different word positions and their effects on intelligibility of non-native speech: All's well that begins well. In O.-S. Bohn and M.J. Munro (eds) *Language Experience in Second Language Speech Learning* (pp. 331–347). Philadelphia, PA: John Benjamins.

Breiner-Sanders, K.E., Lowe, P., Miles, J. and Swender, E. (2000) ACTFL proficiency guidelines: Speaking revised 1999. *Foreign Language Annals* 33 (1), 13–18.

Brown, A. (1991) Functional load and the teaching of pronunciation. In A. Brown (ed.) *Teaching English Pronunciation: A Book of Readings* (pp. 221–224). London: Routledge.

Buckingham, L. (2014) Attitudes to English teachers' accents in the Arabian Gulf. *International Journal of Applied Linguistics* 24, 50–73.

Carey, M.D., Mannel, R.H. and Dunn, P.K. (2011) Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing* 28 (2), 201–219.

Celce-Murcia, M., Brinton, D.M. and Goodwin, J.M. (2010) *Teaching Pronunciation: A Course Book and Reference Guide*. Cambridge: Cambridge University Press.

Chiba, R., Mastuura, H. and Yamamoto, A. (1995) Japanese attitudes toward English accents. *World Englishes* 14 (1), 77–86.

Clarke, C.M. and Garrett, M.F. (2004) Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America* 116 (6), 3647–3658.

Dalton-Puffer, C., Kaltenböck, G. and Smit, U. (1997) Learner attitudes and L2 pronunciation in Austria. *World Englishes* 16 (1), 115–128.

Deterding, D. (2005) Listening to Estuary English in Singapore. *TESOL Quarterly* 39 (3), 425–440.

Eisenchlas, S.A. and Tsurutani, C. (2011) You sound attractive! Perceptions of accented English in a multilingual environment. *Australian Review of Applied Linguistics* 34 (2), 216–236.

Flanigan, B.O. (2006) Different ways of talking in the Buckeye State (Ohio). In W. Wolfram and B. Ward (eds) *American Voices: How Dialects Differ from Coast to Coast* (pp. 118–123). Malden, MA: Blackwell.

Frazer, T.C. (2006) An introduction to Midwest English. In W. Wolfram and B. Ward (eds) *American Voices: How Dialects Differ from Coast to Coast* (pp. 101–105). Malden MA: Blackwell.

Gordon, M.J. (2006) Straight talking from the heartland (Midwest). In W. Wolfram and B. Ward (eds) *American Voices: How Dialects Differ from Coast to Coast* (pp. 106–111). Malden, MA: Blackwell.

Greenwald, A.G., McGhee, D.E. and Schwartz, J.L.K. (1998) Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 74 (6), 1464–1480.

Hay, J., Nolan, A. and Drager, K. (2006a) From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review* 23 (3), 341–379.

Hay, J., Warren, P. and Drager, K. (2006b) Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics* 34 (4), 458–484.
He, D. and Zhang, Q. (2010) Native speaker norms and China English: From the perspective of learners and teachers in China. *TESOL Quarterly* 44 (4), 769–789.
Hu, G. and Lindemann, S. (2009) Stereotypes of Cantonese English, apparent native/non-native status, and their effect on non-native English speakers' perception. *Journal of Multilingual and Multicultural Development* 30 (3), 253–269.
Isaacs, T. (2008) Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *Canadian Modern Language Review* 64 (4), 555–580.
Jenkins, J. (2000) *The Phonology of English as an International Language: New Models, New Norms, New Goals*. Oxford: Oxford University Press.
Kang, O. and Moran, M. (2014) Functional loads of pronunciation features in nonnative speakers' oral assessment. *TESOL Quarterly* 48 (1), 176–187.
Kang, O. and Rubin, D.L. (2009) Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology* 28 (4), 441–456.
Kidd, B. (2010) Superintendent Horne: English teachers cannot have accent [Video, 3 May]. See http://www.azfamily.com/story/28322918/superintendent-horne-english-teachers-cannot-have-accent (accessed 3 May 2015).
Labov, W. (2010) *Principles of Linguistic Change: Cognitive and Cultural Factors* (Vol. 3). Malden, MA: Wiley-Blackwell.
Ladefoged, P. and Johnson, K. (2011) *A Course in Phonetics* (6th edn). Boston, MA: Wadsworth, Cengage Learning.
Lindemann, S. (2003) Koreans, Chinese, or Indians? Attitudes and ideologies about non-native English speakers in the United States. *Journal of Sociolinguistics* 7 (3), 348–364.
Linneman, T.J. (2013) Gender in Jeopardy! Intonation variation on a television game show. *Gender & Society* 27 (1), 82–105.
Lippi-Green, R. (2012) *English with an Accent: Language, Ideology, and Discrimination in the United States* (2nd edn). New York: Routledge.
Litzenberg, J. (2013) An investigation of pre-service English language teacher attitudes towards varieties of English in interaction. Unpublished PhD dissertation, Georgia State University.
McKenzie, R.M. (2008) The role of variety recognition in Japanese university students' attitudes towards English speech varieties. *Journal of Multilingual and Multicultural Development* 29 (2), 139–153.
McKenzie, R.M. (2010) *The Social Psychology of English as a Global Language: Attitudes, Awareness and Identity in the Japanese Context*. London: Springer.
Munro, M.J. and Derwing, T.M. (1995) Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning* 45 (1), 73–97.
Munro, M.J. and Derwing, T.M. (2006) The functional load principle in ESL pronunciation instruction: An exploratory study. *System* 34 (4), 520–531.
Nejjari, W., Gerritsen, M., Van der Haagen, M. and Korzilius, H. (2012) Responses to Dutch accented English. *World Englishes* 31 (2), 248–267.
Niedzielski, N. (1999) The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology* 18 (1), 62–85.
Pantos, A.J. and Perkins, A.W. (2013) Measuring implicit and explicit attitudes toward foreign accented speech. *Journal of Language and Social Psychology* 32 (1), 3–20.
Preston, D.R. (2008) How can you learn a language that isn't there? In J. Przedlacka and K. Dziubalska-Kolaczyk (eds) *English Pronunciation Models: A Changing Scene* (pp. 37–58). Bern: Peter Lang.

Przedlacka, J. (2002) *Estuary English? A Sociophonetic Analysis of Teenage Speech in the Home Counties*. Frankfurt am Main: Peter Lang.
Rajagopalan, K. (2010) The soft ideological underbelly of the notion of intelligibility in discussions about 'World Englishes'. *Applied Linguistics* 31 (3), 465–470.
Rubin, D.L. (1992) Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education* 33 (4), 511–531.
Ryan, E.B. and Bulik, C.M. (1982) Evaluations of middle class and lower class speakers of standard American and German-accented English. *Journal of Language and Social Psychology* 1 (1), 51–61.
Ryan, E.B., Carranza, M.A. and Moffie, R.W. (1977) Reactions toward varying degrees of accentedness in the speech of Spanish-English bilinguals. *Language and Speech* 20, 267–273.
Shockey, L. (2003) *Sound Patterns of Spoken English*. Malden, MA: Wiley-Blackwell.
Simpson, A.P. (2013) Spontaneous speech. In M.J. Jones and R.-A. Knight (eds) *The Bloomsbury Companion to Phonetics* (pp. 155–169). New York: Bloomsbury.
Strand, E.A. (1999) Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology* 18 (1), 86–100.
Thomas, E.R. and Reaser, J. (2004) Delimiting perceptual cues used for the ethnic labeling of African American and European American voices. *Journal of Sociolinguistics* 8 (1), 54–87.
Wolfram, W. and Schilling-Estes, N. (2006) *American English* (2nd edn). Oxford: Blackwell.
Xu, W., Wang, Y. and Case, R.E. (2010) Chinese attitudes towards varieties of English: A pre-Olympic examination. *Language Awareness* 19 (4), 249–260.
Yook, C. and Lindemann, S. (2013) The role of speaker identification in Korean university students' attitudes toward five varieties of English. *Journal of Multilingual and Multicultural Development* 34 (3), 279–296.
Zielinski, B.W. (2008) The listener: No longer the silent partner in reduced intelligibility. *System* 36 (1), 69–84.

12 Teacher-Raters' Assessment of French Lingua Franca Pronunciation

Sara Kennedy, Josée Blanchet and Danielle Guénette

Introduction

The assessment of second language (L2) pronunciation is clearly influenced by raters' understanding of the rated constructs, such as pronunciation accuracy or fluency (see Browne & Fulcher, this volume; Harding, this volume) and by raters' attitudes towards pronunciation (see Lindemann, this volume). This is particularly the case for raters who are also L2 teachers because their understanding and attitudes about pronunciation strongly influence the teaching methods and instructional targets they adopt in both L2 pronunciation teaching and assessment. For example, teachers who prioritize the learning of nativelike suprasegmental production will typically use listening and speaking activities with native speakers producing connected speech that can serve as a model or point of reference. These teachers are likely to favour pronunciation assessments that are tied to a native speaker standard. While teachers themselves may not always be aware that they are adopting or supporting particular values and choices (Harding, this volume), it is important that these values be made explicit. Learners, teachers, raters and other stakeholders can then discuss how these choices contribute to effective teaching, learning, assessment and use of L2 pronunciation. Therefore, the aim of this chapter is to examine the decision making of L2 French teachers in assessing the pronunciation of L2 French learners in the context of Quebec, a Canadian province whose official language is French. We address the following questions: (1) What factors do L2 French teachers associate with particular pronunciation constructs? and (2) Does the use of French as a lingua franca (FLF) factor into teachers' judgements of learners' pronunciation? Our overall objective is to explore teachers' cognitions about assessing L2 French speech and particularly about assessing pronunciation by lingua franca users. We begin

the chapter with an overview of the projected increase in users of FLF, followed by an examination of studies involving the assessment of L2 French pronunciation. Research on raters' decision making in L2 assessment is then presented. Building on the gaps in research identified in previous sections, two research questions are posed, followed by description of the method and presentation of the results ordered by research question. The Discussion section examines possible reasons for raters' focus in their decision making, as a group and as individuals. The chapter concludes with possible implications of the results for language assessment, teaching and research.

French as a Lingua Franca

In this chapter, the use of French as a lingua franca (FLF) is defined as the use of French among speakers who report their first language(s) (L1s) to be other than French (see Seidlhofer, 2013, for broader definitions of lingua franca users). The question of French speakers' L1 status is important because, as for English, the choice of appropriate pronunciation models and norms for teaching, learning and assessment is far from obvious. French is a global language, with 29 countries counting it as the sole or as one of the official languages. An additional 24 countries are official members of the Organisation Internationale de la Francophonie, an international political institution representing countries with ties to French language and culture (Marcoux & Konaté, 2011). It is estimated that in French-speaking Africa, the number of proficient or somewhat proficient speakers of French will double by 2060, comprising over half the habitual speakers of French across the world (Organisation Internationale de la Francophonie, 2010). Because of the current and projected increase in speakers who use FLF, the norms and standards for French language use are increasingly under examination. There have been calls for teachers and researchers of L2 French to acknowledge and value the use of FLF without drawing on a native standard (Johansson & Dervin, 2009). However, compared to research on English as a lingua franca, which includes a growing focus on testing and assessment (e.g. Elder & Harding, 2011) and discussions about the appropriateness of assessment based on a native English norm (see Davies, this volume), the study of FLF is still in its early stages even though its use is projected to increase, especially in African contexts (Marcoux & Harton, 2012).

Assessment of French Pronunciation

In Canada and especially in Quebec, a province where French is the officially mandated language in governmental, commercial and professional contexts, many L2 French teachers and researchers promote the teaching of

varieties indigenous to French Canada (e.g. Auger, 2003; Beaulieu, 2011). However, for research involving *rated assessment* of L2 French pronunciation in Canadian and in other contexts, the norms and criteria used for assessment are often implicit, if described at all. For example, in the Canadian province of Ontario, Knoerr and Weinberg (2005) used one unidentified judge's ratings of controlled production of three elements of pronunciation: the sounds /ã/ and /ʀ/, and intonation contours. No rating scale was described, which invites the question of which pronunciation norm(s) the rater used and what criteria determined whether learners met the norm. In a study in France by Wells (2013), raters were all Francophones who had not spent more than three months outside France and who did not know the L1s of the L2 French speakers. Rating descriptors were taken from the Diplôme d'Etudes en Langue Française (DELF), a standardized test of French proficiency 'awarded by the French Ministry of Education to prove the French-language skills of non-French candidates' (http://www.ciep.fr/en/delf-dalf), with bands corresponding to the Council of Europe's Framework of Reference Levels A1 to B2. The descriptors referred to intelligibility, accent and comprehension problems, but descriptors were not specific about imagined listener characteristics (e.g. L1 or L2 French, amount of previous exposure to L2 French) or particular pronunciation norms. In the United States, Sturm (2013) individually scored read-aloud texts for accurate pronunciation of syllables according to segmental and suprasegmental aspects of speech that had been previously taught. These aspects included syllabification, stress placement and intonation, liaison, and segments. Sturm did not describe which particular pronunciation norms were used to score speech. In a rare exception, Lappin-Fortin and Rye (2014) provided detailed descriptions of how the productions of L2 learners in Ontario were rated, including acceptable variation from standard French.

Description of pronunciation norms and criteria is important in assessment for several reasons. First, segmental and suprasegmental aspects of speech are often different across varieties of native French. For example, in French, the letter <r>, when its quality is not latent (silent), is often spoken as uvular approximant [ʀ], depending on its position in the syllable. This pronunciation is typical of standard French, which does not correspond to any one variety but is typically taught to L2 learners of French (Lyche, 2010). However, the same letter <r> is usually pronounced as apical flap [ɾ] in some (non-standard) varieties of Quebecois French (Detey & Racine, 2012). Learners who are penalized for producing the more easily articulated apical [ɾ] instead of uvular [ʀ] could fairly state that apical [ɾ] is a native variant of <r> (see also Lindemann, this volume; Sewell, this volume). A second reason for clearly outlining pronunciation norms and assessment criteria is because, as mentioned above, the nature of the population of French speakers is changing. L2 speech which may be heavily accented, hard to understand, or unintelligible to a native French speaker may be evaluated quite differently

by a user of FLF, who may have more experience with different L2 accents or may find the relative distance of a speaker's pronunciation from an L1 norm unimportant or irrelevant.

Although there is little research describing the norms and criteria actually used in rating L2 French pronunciation, several researchers have explored French speakers' attitudes and beliefs towards L2 French pronunciation. Ensz (1982) found that, when L1 French speakers from France were presented with L2 speech samples, most preferred L2 speech with little evidence of nonnativeness in terms of pronunciation, grammar and vocabulary use. Drewelow and Theobald (2007) targeted L1 English teaching assistants in North America and a cross-section of native French speakers from France, investigating these listeners' attitudes towards accurate pronunciation by L2 speakers. These researchers observed between-group differences: three-quarters of the teaching assistants thought that native pronunciation was important when interacting with French speakers; however, 88% of L1 French speakers did not believe native pronunciation was important, and 54% thought that learners should not greatly concern themselves with sounding native. Drewelow and Theobald's findings suggest that raters' L1 (i.e. French versus English) is an important factor in their views. Nonetheless, there has been almost no direct exploration of factors underlying raters' decision making in their assessment of L2 French pronunciation, with almost all existing research (described below) focusing on L2s other than French.

Rater Reports as Evidence of Rater Decision Making

There are two complementary approaches to identifying the factors underlying raters' assessment of L2 pronunciation. In this chapter, the two approaches will be referred to as *statistically based* (quantitative) and *cognitions based* (qualitative). A statistically based approach involves examining how test ratings are related to particular rater-, speaker- or speech-specific factors, in order to identify statistically and practically significant relationships. In the cognitions-based approach, raters are asked to explain their ratings decisions, for example, through written or spoken verbal protocols during or after rating. In applied linguistics research, *cognitions* denote participants' thoughts, feelings, impressions and judgements (Cohen, 2013). Verbal protocols eliciting rater cognitions are important in language assessment because, as Isaacs and Thomson (2013) note, even if raters provide similar quantitative assessments they may come to their decisions in qualitatively different ways. Because the current study focuses on the use of verbal protocols to examine rater decisions, research on rater judgements using verbal protocols is reviewed next.

Many verbal protocol studies targeting the assessment of L2 English pronunciation find differences between experienced and inexperienced raters'

comments (see Ballard & Winke, this volume; Saito *et al.*, this volume). In the great majority of verbal protocol studies, experience is operationalized as raters' degree of exposure to L2 accents and not as their amount of rater training. In most cases, no raters have specialized rater training. Teachers of the L2 are typically called *experienced* raters, and participants with no L2 teaching experience or minimal to no training in linguistics are referred to as *inexperienced*. Experienced raters tend to make more comments about suprasegmental elements (Rossiter, 2009) and use more technical vocabulary (Isaacs & Thomson, 2013) compared to inexperienced raters. Raters who are familiar with certain L2 accents have shown awareness of positive bias towards familiar accents and negative bias towards unfamiliar accents, with heritage language speakers (not language teachers) noting strong affective responses towards L1-accented speech from that language (Winke & Gass, 2013).

In verbal protocol studies for assessment of pronunciation, several rating constructs are widely used, typically defined in ways similar to Derwing and Munro (2005). Accentedness, or listeners' perceptions of the degree of difference between a speaker's speech patterns and that of the L1 community, is often explored. Comprehensibility, or listeners' perceptions of their ease of understanding speech, is often investigated together with accentedness to see whether and how raters distinguish between the two constructs, which numerical ratings for L2 English have repeatedly shown to be partially independent (e.g. Munro & Derwing, 1995a, 1995b). Fluency, or listeners' perceptions of the smoothness and rate of speech (Isaacs & Trofimovich, 2011), has been found to be a challenging concept for raters, who have on occasion cited non-temporal aspects of speech (Kennedy *et al.*, 2015). The final construct, communicative effectiveness, has not been investigated to date in verbal protocol studies. In Kennedy and Trofimovich's (2013) study, listeners gave numerical ratings, including for communicative effectiveness, to excerpts of mock job interviews from individual L2 English speakers. Communicative effectiveness can be defined as listeners' perceptions of speakers' ability 'to get [their] message across … to get people's attention … to communicate' (Lehtonen & Karjalainen, 2008, as cited in Kennedy & Trofimovich, 2013: 289). In essence, communicative effectiveness targets how effectively speakers communicate for a specific purpose. In the current study we aimed to give teacher-raters the opportunity to assess not only speakers' accentedness, comprehensibility and fluency, but also the effectiveness with which each speaker communicated (produced or received) messages in paired interaction. Therefore, we also included the construct of communicative effectiveness.[1]

In terms of frequently mentioned linguistic factors underlying raters' judgements in verbal protocol studies, comprehensibility ratings have been explained with reference to accentedness and segmental production (Kennedy *et al.*, 2015), and grammar, vocabulary and fluency measures (Isaacs & Trofimovich, 2012). Wilkerson (2013), one of the few researchers to explore raters' verbal protocols for a non-English L2, analyzed written comments

about accentedness ratings of L1 and L2 German speech. Generally, all raters frequently mentioned speech rate and intonation/rhythm; however, L1 German raters referred to their own comprehension as well as speakers' clarity relatively more frequently, and L1 English raters mentioned speech rate relatively more frequently. Finally, in what seems to be the only study exploring raters' verbal protocols for L2 French, Trottier (2007) investigated five teachers' communicative competence ratings of adult L2 French learners doing interactive role-plays in pairs. When assessing speakers' phonological competence, teachers focused on the use of liaison (linking between words), intonation, segmental production and volume. Teachers' comments about speakers' interactional competence focused on active listening and speakers looking at one another, empathy for weaker partners, and equal distribution of talking time. Notably, when the teachers gave overall communicative competence ratings, they generally assigned equal importance to non-linguistic aspects, such as message coherence and interactional competence.

Taken together, results from verbal protocol studies targeting L2 pronunciation assessment do not show many clearly generalizable patterns of how raters make rating decisions. However, rater experience, usually operationalized as degree of exposure to L2 speech, appears to be linked to different patterns of decision making. Raters who were also L2 teachers tended to prioritize suprasegmental elements and, when rating overall communicative competence, tended to consider non-linguistic aspects to be equally important (e.g. Trottier, 2007). For L2 English pronunciation, some elements frequently mentioned by raters in verbal protocols (e.g. segmental production, grammar, vocabulary for comprehensibility ratings) correspond to relationships found between similar elements and ratings when using a statistically based approach (e.g. Isaacs & Trofimovich, 2012). However, the paucity of research on the assessment of L2 French pronunciation does not allow for conclusions about how and why raters make specific rating decisions. This is important in the case of experienced raters, typically teachers, whose real-world evaluations of L2 speakers usually take place in classrooms. These evaluations both influence and are influenced by the content and type of learning activities that are done in class. If the reasons underlying teachers' evaluation of L2 pronunciation are unclear, then both teachers and learners will be uncertain about the aspects of learners' pronunciation that need to be enhanced.

A lack of clarity would be especially harmful in light of the changing contexts for the use of L2 French. The shifting nature of the French-speaking population means that using a native norm to rate L2 French pronunciation may not always be appropriate (see Sewell, this volume). In West Africa, for example, assessments that penalize nonnative pronunciation, which is widespread in communicative interactions (where many interlocutors are nonnative users of French), may reflect pronunciation norms that are not shared by many in the community. If teachers or other raters are unable to identify and

justify the pronunciation norm(s) they deem as being important (such as standard French versus a local variety) and the presumed future interlocutors for speaking assessments, L2 learners may be taught and/or assessed in ways that do not support their own current or expected future use of French.

The Current Study

Trottier's (2007) focus on paired interaction between L2 French speakers, rather than between an L1 and L2 speaker, is still rare in research on assessment of L2 French. Although Trottier's research took place in the province of Quebec, where French is the official language, the research site of Trottier's study and the current study is Montreal, a city where, according to 2011 census data, over 20% of the population are immigrants (Statistics Canada, n.d.(a)) and 27% of the population report a language other than French as their mother tongue (Statistics Canada, n.d.(b)). In terms of actual FLF use, corpora developers such as Detey and Racine (2012) recognize that the changing demographics of French use means that descriptive data on the use of French by L2 French speakers should be collected. Responding to this research need, Kennedy *et al.* (2015) recently showed that during task-based interactions between pairs of L2 French speakers, pronunciation accounted for 18% of identified comprehension problems between the pairs (as revealed through stimulated recalls of speaker interactions). The elements most frequently linked to comprehension problems were segments, particularly consonant production. However, this is the only published study on specific pronunciation elements linked to communication difficulties in FLF use. It is important to investigate how not only student peers but also teachers react to pronunciation from speakers using FLF. Do teachers tend to adopt an L1 French norm of pronunciation when assessing FLF users, or does the context of interaction (between L2 French speakers) elicit a more multi-faceted approach in teachers' assessment? Teachers' rating decisions have implications for the nature of instruction and feedback on pronunciation that is given in the classroom. For example, over 99% of the population of Mauritius speak either Mauritian Creole or Bhojpuri as L1s, but many Mauritians use French for business interactions and local media production and consumption (Chiba, 2006). In this context, then, those Mauritian teachers of French who penalize 'non-standard' French pronunciation may be requiring a native norm that is less relevant for local interlocutors' typical interactions in French.

Therefore, it is important that the reasons for teacher-raters' decisions are explicit and open to examination by colleagues, supervisors and students. With the exception of Trottier (2007), no published research explores the factors underlying teachers' rating decisions for L2 French pronunciation. There is almost no existing knowledge base about how teachers of L2 French interpret constructs related to pronunciation assessment. Given this lack of

information, we chose to conduct a case study of four experienced teachers (henceforth, teacher-raters) to explore how they interpreted particular constructs and what factors they mentioned while rating for those constructs. In essence, we focused not on the ratings themselves (which will not be discussed here), and were predominantly interested in teacher-raters' *reasons* underlying their ratings. In this study, teacher-raters were classified as experienced because of their extensive familiarity with hearing, teaching and assessing L2 French speech in classrooms. To avoid prematurely shaping the findings, particular factors were not preselected for investigation. Therefore, all of the teacher-raters' utterances were treated as potentially relevant to their decision making. As a case study, the findings below are not meant to be universally generalized to other contexts (VanWynsberghe & Khan, 2007), but to be seen as an early step in the construction of a knowledge base about how teachers of L2 French make rating decisions about pronunciation. We asked the following research questions:

(1) What factors do experienced L2 teachers of French associate with the constructs of accentedness, comprehensibility, fluidity (fluency) and communicative effectiveness in L2 French speech?
(2) How does FLF use relate to teachers' judgements of learners' pronunciation?

Methodology

Teacher-raters

Four teacher-raters participated in the study: Raymond, Sandrine, Klara and Julie (all pseudonyms). Sandrine, Julie and Raymond had MA- or PhD-level degrees and taught in the same for-credit certificate programme in French as a second language at a French-language university in Montreal, Quebec, Canada. Klara had an MA degree and was an instructor in a similar programme in an English language university in the same city. All teacher-raters had a minimum of a decade of experience teaching L2 French in Quebec and, for Klara and Sandrine, in other countries, and all had experience teaching L2 French pronunciation (see Table 12.1). Three of the four teacher-raters taught courses in the certificate programme from which speakers had been recruited, but no one reported familiarity with any speaker.

FLF speakers

The initial participant pool comprised 18 students who were L2 speakers of French at intermediate and advanced levels enrolled in the teacher-raters' L2 French certificate programme; they had volunteered to participate in a study on spoken interaction (Guénette *et al.*, in press; Kennedy *et al.*, 2015).

Table 12.1 Teacher-raters' reported language learning, teaching and professional background

	Raymond	*Sandrine*	*Klara*	*Julie*
L1	*French*	*French*	*Russian*	*French*
Accent (self-reported)	Native standard Québecois	Native France	Quasi-native; mix of European and Québecois influences	Native Québecois
Other spoken languages	English	Spanish, English	French, English	English
Academic background	PhD (Linguistics, phonology specialization, coursework completed); MA (Linguistics)	PhD (Applied Linguistics, specialization – Phonetics, first year); MA (Education, teaching French as a foreign language)	MA (Applied Linguistics); BA (Linguistics, second language teaching)	MA (Applied Linguistics, phonetics)
L2 French teaching experience (years)	11	25	18	20
Percentage teaching time (L2 pronunciation)	80	50	50	25
Official assessment qualifications	None	None	Certified examiner for a standardized language test	None

The speakers were filmed in pairs, interacting with one another in French while completing a map task (described below). When possible, the pairs included students of different L1s, but students' availability for filming was the deciding factor in the composition of pairs. Recordings of three pairs were selected for the current study to include a range of L1s and L2 proficiency levels so that teacher-raters could react to a range of L2 speech and interactions (see Table 12.2). Two pairs had speakers from different L1s (Russian-Spanish and Ukrainian-Chinese), and one pair had same-L1 speakers (Chinese). Skill levels in French, as determined by self-rating, varied across pairs.

Table 12.2 Information on speakers

Pair	*L1*	*Names*	*Sex*	*Self-assessment*[a]		*Courses completed*
				Speak	*Listen*	
1	Russian	Darya	F	6	6	Oral communication; oral comprehension; grammar and writing; reading (all Advanced I level)
	Spanish	Carmen	F	7	6	Oral communication; oral comprehension; grammar and writing; reading (all Advanced I level)
2	Chinese	Ying	F	2	2	Oral communication; oral comprehension; grammar and writing; reading (all Advanced I level)
	Chinese	Hua	M	4	4	Oral communication (level unknown)
3	Russian	Alexa	F	5	7	Four-skills (Intermediate level)
	Chinese	Chen	F	5	5	Oral communication; grammar and writing; (all Advanced I level)

Note: [a]On a scale of 1–9 (1 = very weak, 9 = very strong).

Task and measures

Speakers completed an information-gap map task, following Lindemann (2002). The goal of the task was for each speaker to exchange information in order to end up with similar maps. One speaker had a version of the map with 10 landmarks and no route while the other speaker had a map containing six landmarks (four were missing) and a route. The speakers had a maximum of seven minutes to exchange information verbally in French, without seeing each other's map, so that the finished maps would be identical. Each speaker was assessed by teacher-raters on five-point Likert-type scales for four constructs, with short descriptors in French only at endpoints.[2] Five-point scales were selected over scales featuring nine points because there is no compelling evidence to prefer the latter scale (Isaacs & Thomson, 2013), and it was hoped that the five-point scale would allow teacher-raters to focus more on qualitative rather than quantitative assessment of the speech. We analyzed data from the four constructs relating to L2 pronunciation: accentedness (1 = 'very strong', 5 = 'very weak'), comprehensibility (1 = 'very

difficult to understand', 5 = 'very easy to understand'), fluidity (1 = 'not at all fluid', 5 = 'very fluid'), and communicative effectiveness (1 = 'ineffective', 5 = 'very effective').[3]

Rating procedure

Teacher-raters, who reported having normal hearing, completed an individual rating and verbal protocol session in French with the second author (henceforth, the researcher), a native French speaker and a current or former colleague of each rater. Sessions lasted between 90 and 110 minutes. On average, raters spent 22 minutes rating and commenting on one pair of speakers. The verbal protocol procedure described below was based on Ducasse and Brown (2009), with the additional participation of the researcher. The teacher-raters were informed that the aim of the study was to better understand evaluations of L2 French speakers' language skills. They then reviewed the paper-based rating scales and the two versions of the map task that the speakers had been presented with. Teacher-raters were instructed to view each video-recorded interaction once without interruption on a desktop computer using external speakers; they could then navigate within the video while rating speakers and orally comment on their ratings. They could circle two numbers if they wished to add a half point to a score. The constructs were not explicitly defined so as not to influence the teacher-raters; however, the researcher provided basic answers if questions arose. Teacher-raters were told that communicative effectiveness was not necessarily linked to successful completion of the map task. A speaker could be deemed effective even if he or she did not complete the task or did not complete it accurately.

Teacher-raters wore a lapel microphone to record their speech onto the same desktop computer. Following a practice task to familiarize them with the procedure, they viewed the first of three seven-minute video recordings in unique randomized orders, completing the rating scales and orally commenting on their ratings before moving to the next pair. Speakers' faces were digitally blurred to safeguard confidentiality. The researcher asked questions when it was necessary to elicit information or comments on constructs that had not yet been discussed. The researcher also stopped the recording if she felt the teacher-rater had something to convey, had remained silent for a long time, or had modified a rating (in order to find out what motivated the changes). Four weeks or more after the rating session, teacher-raters completed an emailed 25-item questionnaire in French about language learning history, academic background, teaching experience and attitudes. Items comprised multiple-choice, short-answer and Likert-type scales.

Data analysis

The teacher-raters' comments were transcribed in French by a research assistant and the second and third authors, all native French speakers, using

broad orthographic transcription. Transcripts were coded by the first two authors (a high-intermediate and a native speaker of French, respectively) working both separately and in multiple joint sessions to refine and apply the codes. All final coding decisions were made by consensus. Codes were developed using a combination of a priori coding and empirical codes drawn from the content of the transcripts (following Gibson & Brown, 2009). The *a priori* codes followed the rated constructs and were used to identify comments relating to the research questions. The codes were attributed either when the rater explicitly linked the comment to a specific construct or when the discourse preceding or following the comment demonstrated a focus on a specific construct, as shown for comprehensibility in Excerpt 11.1 (translated from French).

Excerpt 11.1

Researcher: So here, for comprehensibility, for Alexa, it's not so much about segmental substitutions but rather about prosody? Or are there other factors at play?

Rater: Well, no, there are other factors! It's as a whole, really: vocabulary, syntax. The vocabulary is also very limited.

The empirical codes were derived from repeated reading of the transcripts and refining of themes emerging from the comments (see Appendix to this chapter). All empirical codes were tallied and analyzed by construct, with each rater's comments tallied separately within each construct. Some of the codes refer to linguistic, behavioural or contextual factors, while other codes refer to norms and processes for rating or to raters' perceptions of the speakers. A given utterance could be tagged with multiple codes and a code could be linked to additional codes (e.g. comprehensibility linked to communicative effectiveness). Beginnings and endings of utterances were identified as comments which could in their entirety be categorized by one of the six a priori codes or as comments which in their entirety could not be linked to any of such codes. The unlinked set of comments is not presented in the results because it does not relate to the two research questions. Comments making global evaluations, such as 'They're really good', were also not included in the final tally of results.

Results

Research Question 1

The first research question asked what factors L2 French teachers associated with the constructs of accentedness, comprehensibility, fluidity and communicative effectiveness in L2 French speech. A table showing the frequency of all factors mentioned for each construct appears in the Appendix.

Below, we present, for each construct in turn, the four most frequently mentioned factors overall, with their relative percentages given in parentheses. In addition, the first and second most frequently mentioned factors for each of the four teacher-raters are presented for each construct.

Accentedness

The 104 comments relating to accentedness generally related to salience, speakers' L1s and segments. Overall, the salience of speakers' accents was noted almost exclusively in discussions of accentedness (18% of all comments) but not in discussions of other constructs. References to speakers' L1s were also regularly made (16%), as in 'We really hear the Chinese behind that', with somewhat fewer mentions of segmental production (12%). Teacher-raters also compared the perceived level of accentedness between different speakers (10%), as in 'It's clear that [her] accent is stronger'. Individually, the teacher-raters showed similar patterns. The two factors most frequently mentioned by Raymond were the speakers' L1s and the perceived salience of accents, while Sandrine most often noted segmental production and salience. Klara commented most frequently on salience, followed by equally frequent mentions of speakers' L1s, and segmental and suprasegmental production. Julie most often commented on speakers' L1 but, interestingly, just as many of her comments focused on task type and its contribution to her ratings and on syntax and morphology, as in: 'If you say to me *del*, *al*, at some point, if I hear Spanish, that affects accent […] it's in that sense.'

Comprehensibility

The 188 comments on comprehensibility encompassed a wide variety of factors mentioned by teacher-raters. Overall, the factor most frequently mentioned related to whether a native speaker would understand the speaker (13%). Pronunciation in general (with no specific aspect mentioned) was noted almost as frequently (12%). Speakers' knowledge and use of lexis was also remarked on (10%), such as 'It's hard to understand, he has a very limited vocabulary.' Somewhat less frequent were comments where teacher-raters considered to what extent the nonnative speaker pairs understood each other (8%). The concerns of individual teacher-raters did not always reflect the overall pattern. Raymond most often commented on speakers' general pronunciation and on their comprehensibility to a native speaker. Interestingly, for two different pairs, Raymond mentioned links between comprehensibility and accentedness, although for a speaker in one of the same pairs he noted that other factors played a greater role in comprehensibility than accentedness, as exemplified in the following: 'So I find that accent does not play the main role in comprehensibility difficulties …'. Sandrine made equally frequent mention of speakers' inaccurate production of segments and of native speakers' understanding. Her second most frequent comments were about her difficulty in evaluating the comprehensibility of one reticent speaker and about how

speakers compared to one another in their comprehensibility. Klara, a non-native speaker of French, commented most often on speakers' understanding of each other as well as on their segmental production. Her second most frequent comments were about native speakers' understanding of the L2 speech and about lexical aspects, with the assumption that the speaker pronounced the word in a non-target fashion because he had only an approximate knowledge of the word, as shown in 'It's hard to understand, he has a very limited vocabulary.' Julie commented most frequently on native speakers' understanding of the speech and she also noted the contribution of task type to comprehensibility. She stated that a goal-driven interaction under time pressure is very different from a normal conversation and that the speakers may appear less comprehensible in this situation than they would normally be.

Fluidity

In the 103 comments for fluidity, the most frequent factor mentioned was continuity (16%), with lexis also regularly commented on (14%), sometimes together with continuity, as in 'I asked myself, Do they hesitate much? Are they grasping for words or not?' Teacher-raters also compared speakers' fluidity (13%). Finally, they noted the challenge of assessing fluidity, particularly for the pair with one diffident speaker (10%). Most individual teacher-raters also frequently mentioned these factors. Raymond's two most frequent factors mentioned were continuity and lexis, as shown in 'There is a bit of searching for the exact term but they don't search long, which is why they're doing well on fluidity.' Sandrine often commented on speakers' continuity and speech rate, as well as their knowledge and use of lexis. Klara frequently compared speakers' fluidity and also commented on her ability to judge fluidity. For Julie, task type was again mentioned frequently because the task forced speakers to be brief and made it difficult to assess fluidity; she also commented on speakers' participation in the interaction, especially with regard to the reluctant speaker.

Communicative effectiveness

The highest number of comments (243) was made about this construct, and most comments were not specifically language related. Speakers' use of communication strategies was by far the most frequent factor mentioned (24%), followed by speakers' participation in the interaction (16%). Sometimes these factors were linked by teacher-raters, as in 'She has more vocabulary and much more efficient strategies, I mean that she immediately takes charge and proposes a strategy for action.' Teacher-raters also mentioned difficulty in assessing communicative effectiveness, specifically for the speaker who spoke very little (10%). Finally, teacher-raters commented equally frequently about individual speakers' oral comprehension and about speakers' comparative communicative effectiveness (9%). Individual teacher-raters closely reflected the overall patterns, with Raymond mentioning communication strategy use and comparisons between speakers most frequently,

followed by speaker's participation in the interaction. Sandrine and Julie also commented most frequently on communication strategies and speakers' participation, while Klara spoke most often about communication strategies and speakers' use and knowledge of lexis.

Excerpt 11.2

Klara: Is it a lack of vocabulary or a lack of comprehension strategies? We have to consider developing strategies, here. Often students are blocked because they are focusing on the elements that they don't understand. Strategies should be worked on more: how to construct meaning from the elements we do understand and from context.

Summary: Research Question 1

Teacher-raters showed clear differences in the factors they discussed for each construct. For accentedness, the salience of the accent, speakers' L1s and segments were often noted. For comprehensibility, the comprehension of native speakers, pronunciation, lexis, and mutual comprehension within nonnative pairs were often considered, although individual raters sometimes prioritized different factors. In rating fluidity, teacher-raters paid attention to performance-based factors such as continuity and use of lexis, but also discussed the process of coming to a rating decision, whether by comparing speakers or noting the difficulty in rating a given speaker. Teacher-raters' concerns about communicative effectiveness were mostly clear cut, with many comments tied to use of communication strategies and speakers' participation in the interactions. Communicative effectiveness was also shown to be one of the constructs in which the use of FLF was highlighted by raters, as shown below.

Research Question 2

The second research question centred on how FLF use factored into teacher-raters' judgements of learners' pronunciation. The factors that are particularly relevant to this question are the use of a native speaker norm in rating and references to assessing the nonnative speaker pairs in relation to their own interactions. The overwhelming majority of comments on these factors (95%) appeared in ratings of speakers' comprehensibility and communicative effectiveness, presented below.

Comprehensibility

When analyzed as a whole, teacher-raters' comments showed clear preferences for drawing on a native speaker norm for comprehensibility (62%) over referring to the nonnative speakers' understanding of each other (38%). Statements such as 'We lose certain sounds, certain vowels, certain words. I have to recap … it takes me a little while, there is a little gap before

I recognize the word that was said or pronounced', were more frequent than remarks like 'They understand each other.' However, individual teacher-raters varied in their emphasis on a native speaker norm. Raymond mentioned several times that the two speakers understood each other, but his comments that implicitly or explicitly assumed an L1 French listener were almost twice as frequent. In rating comprehensibility for all speaker pairs, Sandrine made explicit mention of her standard of comprehension by L1 Francophones: 'For the comprehensibility factor, I refer to a francophone listener when judging.' For Klara, the lingua franca aspect was often noted when judging comprehensibility, and was twice as frequent as Klara's mention of an L1 French norm. However, she also sometimes adopted an L1 French speaker's point of view. Julie included both lingua franca and native speaker norm perspectives. When talking about one pair, for example, she said, 'They understand each other and I understand them', with both criteria considered to be conditions for good performance in comprehensibility. However, her mentions of a native norm were somewhat more frequent than comments about speakers' mutual understanding.

Communicative effectiveness

Although less clear cut, teacher-raters' discussions of communicative effectiveness demonstrated a tendency for speakers' mutual understanding to be given more attention than the use of a native speaker norm (with a ratio of 7:3 comments, respectively). Nevertheless, individual teacher-raters again differed in their focus on each factor. Raymond referred to speakers' understanding of each other almost as often as he did to his own, while Sandrine did not mention native speaker norms at all but referred only to speakers' mutual understanding, as in 'In communicative effectiveness I'm taking into account communication breakdowns.' Klara seldom mentioned comprehension, but when she did, she referred to speakers' understanding of each other: 'She has difficulty pronouncing words but her partner understands her anyway.' Julie did not mention either a native speaker norm or speakers' mutual understanding in her comments on communicative effectiveness.

Summary: Research Question 2

Considerations of FLF use tended to be mentioned more frequently in the rating of communicative effectiveness than comprehensibility. However, even when rating comprehensibility, most teacher-raters did make some comments about speakers' understanding of one another. Teacher-raters also differed in the seeming importance they gave to the use of FLF. Some raters explicitly held to a native speaker norm (Sandrine for comprehensibility). Others mentioned both a native speaker norm and speakers' mutual understanding, but for particular constructs referred more often to an L1 speaker's point of view (Julie and Raymond for comprehensibility), or to the use of FLF (Klara for comprehensibility).

Discussion

Hardly any research has been conducted on the assessment of L2 French pronunciation, so there are few grounds on which to compare current findings to previous findings for L2 French. With respect to findings from verbal protocol studies for other L2s, there are some similarities but also notable differences (see Table 12.3). As in some L2 English studies, as a group, teacher-raters in the current study linked accentedness and segmental production and linked comprehensibility to pronunciation and lexical knowledge and its use. Additionally, continuity and pausing were mentioned in the assessment of fluency. Although similarities between current and previous findings are interesting, results for the assessment of L2 English pronunciation may have minimal relevance for the assessment of L2 French pronunciation, especially regarding the linguistic factors linked to different constructs. The current study was conducted with L2 French speakers evaluated by

Table 12.3 Frequently mentioned factors in current and past verbal protocol studies

Construct	*Current study – L2 French (paired)*	*L2 English*[a]
Accentedness	• salience • L1 influence • segmental production • comparison of speakers	• segmental production
Comprehensibility	• comprehended by native speaker • general pronunciation • lexis • speakers' mutual understanding	• accentedness • segmental production • grammar • fluency • lexis • listener familiarity • irritability • word-based vs. discourse-based understanding
Fluidity	• continuity • lexis • comparison of speakers • ability to evaluate	• pausing • rhythm • self-repetition • speech rate • self-corrections
Communicative effectiveness	• communication strategies • participation in interaction • ability to evaluate • oral comprehension and comparison of speakers	• no data

Source: [a]Isaacs and Thomson (2013), Isaacs and Trofimovich (2012), Isaacs *et al.* (2015), Kennedy *et al.* (2015) and Rossiter (2009).

experienced teachers of L2 French. Because the raters had extensive knowledge both of French as a linguistic system and the teaching of L2 French, it is worth exploring more deeply some novel findings of the current study that might have implications for the assessment of L2 French by experienced teacher-raters.

Raters' focus as a group and as individuals

As noted above, the preoccupations of individual teacher-raters were sometimes very similar for a given construct, but for other constructs clear differences were shown. Because this was an exploratory study, all factors that were mentioned in the process of assessment were analyzed. These included comments not only about linguistic features, but about speakers' backgrounds, the assessment process and other areas.

Accentedness

In terms of accentedness, teacher-raters did not link speakers' level of accentedness to their level of comprehensibility, but made explicit distinctions between the two constructs. This distinction could be tied to the teacher-raters' academic training and teaching experience with French pronunciation, as inexperienced raters in previous research have discussed comprehensibility in terms of accentedness (e.g. Kennedy *et al.*, 2015). The frequent comments about speakers' perceived L1s suggests teacher-raters' tendency to consider speakers' L2 accents not in the abstract, but as a product of speakers' L1 backgrounds. Teacher-raters' attention to the perceived differences in level of accentedness between speakers is a telling indication of how these teacher-raters determined a speaker's level of accentedness in this assessment context; part of the process was comparing speakers to each other, which suggests that for the assessment of some constructs, the score that raters give to Speaker A may be influenced by the performance of other speakers who were assessed by the raters before, during or after Speaker A's performance/turn.

Some factors mentioned in rating accentedness were specific to individual teacher-raters. For example, Julie included speakers' syntax and morphology and task type. For Julie, incomplete sentences and missing words contributed to the overall impression she formed of speaker accent. In addition, because preposition–article contractions occur in both French (e.g. *au* = *à* + *le*) and Spanish (e.g. *al* = *a* + *el*), Julie recognized the grammatical accuracy of an L1 Spanish speaker's systematic use of Spanish contractions; however, she interpreted the surface-level Spanish form as a particularly salient demonstration of L2 accent. This is an example of how the relationship between learners' L1 and L2 (in this case, Spanish and French) can affect the linguistic factors (e.g. morphology) mentioned by raters (Sewell, this

volume). Regarding task type, Julie noted that the map task induced speakers to structure their interactions in particular ways, which affected the length of their utterances and thus Julie's ability to assess speakers' suprasegmental production.

Comprehensibility

Many of the factors frequently mentioned for comprehensibility are addressed in the discussion of FLF below. In terms of individual teacher-raters, Julie again noted the influence of task type, suggesting that the goal-driven and time-dependent nature of the task may have put inordinate pressure on some speakers, leading to less comprehensible speech. Rater awareness of the influence of task type on L2 pronunciation has been recently reported by Hayes-Harb and Hacking (2015), and task effects have been found to be important for comprehensibility ratings of L2 English in correlational research (Crowther *et al.*, 2015). Raymond and Sandrine also mentioned the difficulty of rating a reticent speaker's comprehensibility, demonstrating their understanding that assessments based on incomplete or insufficient information could be criticized as unfair or unjustified.

Fluidity

As a group, the teacher-raters again showed their inclination to compare speakers to one another in their assessment of fluidity, which could be criticized on the grounds that, even in their L1s, speakers differ in the fluidity and rate of their speech (Segalowitz, 2010). However, teacher-raters showed a sophisticated level of differentiating between different aspects of fluidity, commenting on speakers' lexical knowledge and retrieval, rather than simply on surface-level phenomena such as hesitations. Individually, Klara noted that her fluidity ratings suffered from lack of information about the context, with no details about learning objectives or future L2 use. Klara's comments reflect an assumption that, for a rating to be done well, the rater must have some idea of what is or will be expected of the speaker.

Communicative effectiveness

Communicative effectiveness as a construct has not typically been included in L2 pronunciation research. Therefore, the factors mentioned in relation to communicative effectiveness are novel findings. Teacher-raters had been instructed that communicative effectiveness was not necessarily linked to successful completion of the task, but this was the only guidance they had received. In explaining their ratings, all raters put speakers' use of communication strategies and their participation in the interactions at the forefront. Speakers' comprehension of their partners was also important for some raters. These results reflect Trottier's (2007) findings about the

importance of active listening and the equal distribution of talking time in rating interactive competence for L2 French speakers. In the current study, the linguistic aspects of speech, such as pronunciation and vocabulary, were rarely mentioned in communicative effectiveness ratings. It is unclear whether linguistic aspects would have been downplayed if raters had not also assessed comprehensibility. For most raters, comprehensibility ratings were linked to phonological and lexical elements; raters seemed to address speakers' strategic and interactive behaviour with communicative effectiveness ratings. In this study, therefore, it seems that raters considered communicative effectiveness to be somewhat independent of the linguistic form of L2 speech.[4] The steps taken by individual speakers to successfully exchange information seemed to be most important.

To summarize, then, experienced teacher-raters commented on 'surface' elements of language and speech, such as segmentals and suprasegmentals, lexis, morphosyntax and speech rater, but also made comments that were not solely form based (see Harding, this volume). In their verbal reports, raters mentioned speakers' cognitive processes (e.g. access to lexis), the nature of the interaction (e.g. use of communication strategies), and other influences on speakers' production (e.g. task). This wide-ranging focus may be likely to have stemmed from raters' extensive experience with teaching and evaluating L2 French in classroom contexts, which may have sensitized them to the diverse aspects (such as task type) that can affect L2 speech and its assessment.

Raters' perceptions of French as a lingua franca

All raters demonstrated attention to the degree of mutual understanding between each pair of speakers. However, raters differed in the extent to which they assumed a native French speaker when assessing comprehensibility. Raymond and Sandrine showed clear preferences for a native speaker norm, while Julie and Klara adopted both native speaker and lingua franca perspectives when rating, with Klara making frequent mention of pairs' mutual understanding. On the other hand, when raters assessed communicative effectiveness, most paid careful attention to understanding between speakers, suggesting a focus on mutual understanding of lingua franca speakers (see also Isaacs, 2013).

Raymond and Sandrine, the two raters who relied most heavily on the native norm when rating comprehensibility, had the most advanced formal training in phonology and phonetics and taught at least 50% of their yearly course load in that area. In addition, in the rater questionnaire completed after the rating session, these two raters showed strong agreement with statements about the importance of intelligibility, comprehensibility and nativelike accent for L2 speakers hoping to integrate into the workplace and society in general. Klara, the rater who seemed to show a preference for rating speakers by their mutual understanding, was different from the other

raters in that she did not grow up in a French-medium environment, but had first learned French through instruction and moved to a French-medium environment as an adult. This early experience as a learner of French could account for her acceptance of the (lingua franca) speakers' perspectives in assessing comprehensibility. In the rater questionnaire, Klara showed the lowest level of agreement out of the four raters about the importance of intelligibility, comprehensibility and nativelike accent for the workplace and societal integration. Her disagreement with the statement 'A native or nativelike accent in French helps immigrants to integrate into Quebec's society' may reflect an awareness that other characteristics, such as curiosity or tolerance of difference, may be more important for integration (see Zhang & Elder, 2014, for other research on differences between native and nonnative raters). It would appear, then, that individual differences such as raters' formal training in phonetics and phonology, the types of courses they taught, and the nature of their initial learning environment could be important for raters when assessing the comprehensibility of speakers using FLF (see Isaacs & Trofimovich, 2011; Kennedy & Trofimovich, 2013; and Sewell, this volume, for additional research on rating differences based on programme of study).

Limitations and Conclusion

The current study was a case study; therefore, the findings are particular to the research context and are not meant to be generalized. Although French is the official language of the province of Quebec, English is often commonly used in daily life in the city of Montreal. In settings where one language is clearly dominant in daily life, whether French or another language, teacher-raters may explain ratings decisions quite differently. In addition, the task and scales used for ratings were developed for research and were not drawn from existing courses or assessment instruments. In future, teacher-raters' process and rationale for ratings should be explored with tasks and rubrics that are similar to those used in their classrooms.

Despite these limitations, the findings from this exploratory case study offer several conclusions about the assessment of L2 French pronunciation by teacher-raters with considerable classroom experience. All understood the difference between a speaker's accentedness and his/her comprehensibility, and were able to rate the two constructs separately. The raters regularly mentioned the influence of formal aspects of speech on their rating of constructs, but also took into account cognitive, pragmatic and task-based information, showing their awareness of different factors that may affect both L2 speech and raters' assessment of it. In terms of FLF, while making rating decisions for comprehensibility and communicative effectiveness, all raters noted the use of FLF and speakers' mutual understanding. Teacher-raters made deliberate choices about the weight they gave to L1 pronunciation

norms and to speakers' ability to understand one another. The deliberateness of raters' choices, with little evidence of hesitation or uncertainty, suggests that these raters had stable understandings of each construct. However, raters did not always conceive of a given construct (e.g. accentedness or comprehensibility) in the same way. This lack of consensus will be addressed in the next section.

Implications for Assessment, Teaching and Research

Although the teacher-raters in this study mentioned a common set of factors while rating, they also noted other factors that were individual to particular raters. These disparities can lead to problems for teaching and assessment, especially if teachers, raters and learners have different conceptions of how and why a particular construct is assessed in a particular way (Douglas, 1994; Isaacs & Thomson, 2013). All parties would benefit from open and regular discussion of what teacher-raters are considering in their classroom assessments – especially in L2 pronunciation, an area in which a variety of pedagogical approaches and ideological stances have recently come into contact (e.g. valuing 'standard' accents versus language as a lingua franca). It is important that future research explore not only what speech elements raters attend to in their ratings, but also what they *believe* is important for L2 communication. When the connection is made explicit, then raters, teachers, learners and other stakeholders can have more informed discussions about how to promote fair, principled, consistent and effective assessment of L2 pronunciation.

The current study is one of only a few that qualitatively explore pronunciation rating decisions for a language other than English. As such, it highlights factors, such as morphology, which may be important for teacher-raters because of the linguistic structure of French itself. Because the study was not limited to analysis of linguistic factors, the study also showed teacher-raters' attention to non-linguistic factors such as task type and the rating process. It is hoped that future research on rating decisions for L2 pronunciation will see an increasing number of studies on languages other than English, incorporating classroom teachers and, ideally, classroom tasks and rating rubrics. Teachers perform the earliest and most frequent assessment of classroom language learners, and so must assume a central role in research on L2 pronunciation assessment.

Acknowledgements

We gratefully acknowledge the participation of the four teacher-raters. This chapter is dedicated to Danielle Guénette, a wonderful researcher,

colleague and friend who fully contributed even while confronting a terminal illness. We miss her very much.

Notes

(1) In the current study, the construct of communicative effectiveness has a surface resemblance to the construct of interactional competence (e.g. He & Young, 1998). However, there are important differences between the two. First, interactional competence is conceived a priori as jointly constructed by interlocutors and not separable from the interactive practice in which it is observed (He & Young, 1998). Communicative effectiveness, however, is envisaged as a construct which, although potentially influenced by interlocutors and other contextual factors, resides in individual speakers. The second difference is that interactional competence relates to the 'skilful use of resources' in co-constructing some discursive act (He & Young, 1998: 7), while communicative effectiveness is concerned principally with the effective communication of meaning, which could involve and be shaped by interaction between interlocutors, but could also be assessed in monologic speech.

(2) All translations from the original French were done by Josée Blanchet, a balanced French-English bilingual with a degree in translation.

(3) Communicative effectiveness was assessed for individual speakers and not in pairs because the construct is considered by the authors to reside primarily in the individual speaker (see Rater reports section for more details).

(4) Interestingly, teacher-raters seemed to have no difficulty in assessing the communicative effectiveness of individual speakers, frequently noting for one pair the minimal participation by one speaker and the strenuous efforts made by the other. This echoes findings from previous L2 research on interactional patterns in collaborative speaking tasks related to the occurrence of unbalanced interactions (Galaczi, 2008; Storch, 2002) and to interlocutors' occasionally mismatched perceptions of the relative effort required to understand one another (Isaacs, 2013).

References

Auger, J. (2003) Le français au Québec à l'aube du vingt et unième siècle. *The French Review* 77 (1), 86–100.

Beaulieu, S. (2011) Norme pédagogique et infirmières bilingues en milieu francophone minoritaire. *Revue Canadienne des Langues Vivantes* 67 (4), 508–535.

Chiba, E. (2006) English use in Mauritius. See http://homes.chass.utoronto.ca/ ~ cpercy/courses/6362-chiba.htm.

Cohen, A.D. (2013) Verbal report. In C.A. Chapelle (ed.) *Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell.

Crowther, D., Trofimovich, P., Isaacs, T. and Saito, K. (2015) Does a speaking task affect second language comprehensibility? *Modern Language Journal* 99 (1), 80–95.

Derwing, T.M. and Munro, M.J. (2005) Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly* 39 (3), 379–397.

Detey, S. and Racine, I. (2012) Les apprenant de français face aux normes de prononciation: Quelle(s) entrée(s) pour quelle(s) sortie(s)? *Revue Française de Linguistique Appliquée* 17 (1), 81–96.

Douglas, D. (1994) Quantity and quality in speaking test performance. *Language Testing* 11 (2), 125–144.

Drewelow, I. and Theobald, A. (2007) A comparison of the attitudes of learners, instructors, and native French speakers about the pronunciation of French: An exploratory study. *Foreign Language Annals* 40 (3), 491–520.

Ducasse, A.M. and Brown, A. (2009) Assessing paired orals: Raters' orientation to interaction. *Language Testing* 26 (3), 423–443.

Elder, C. and Harding, L. (2011) Language testing and English as an international language constraints and contributions. *Australian Review of Applied Linguistics* 31 (3), 34.1–34.11.

Ensz, K.Y. (1982) French attitudes toward typical speech errors of American speakers of French. *Modern Language Journal* 66 (2), 133–139.

Galaczi, E.D. (2008) Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly* 5 (2), 89–119.

Gibson, W. and Brown, A. (2009) *Working with Qualitative Data*. London: Sage.

Guénette, D., Kennedy, S., Allard, S. and Murphy, J. (in press) Interactions verbales et résolution de malentendus en français L2 entre locuteurs de L1 commune et différente: Une étude de cas. *Language, Interaction, and Acquisition.*

Hayes-Harb, R. and Hacking, J.F. (2015) Beyond rating data: What do listeners believe underlies their accentedness judgments? *Journal of Second Language Pronunciation* 1 (1), 43–62.

He, A.W. and Young, R. (1998) Language proficiency interviews: A discourse approach. In R. Young and A.W. He (eds) *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency.* Amsterdam: John Benjamins.

Isaacs, T. (2013) International engineering graduate students' interactional patterns on a paired speaking test: Interlocutors' perspectives. In K. McDonough and A. Mackey (eds) *Second Language Interaction in Diverse Educational Settings* (pp. 227–246). Amsterdam: John Benjamins.

Isaacs, T. and Thomson, R.I. (2013) Rater experience, rating scale length, and judgements of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly* 10, 135–159.

Isaacs, T. and Trofimovich, P. (2011) Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics* 32 (1), 113–140.

Isaacs, T. and Trofimovich, P. (2012) Deconstructing comprehensibility. *Studies in Second Language Acquisition* 34 (3), 475–505.

Isaacs, T., Trofimovich, P., Yu, G. and Chereau, B.M. (2015) Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS Research Reports Online* 4, 1–48.

Johansson, M. and Dervin, F. (2009) Cercles francophone et français lingua franca: Pour une francophonie liquide. *International Journal of Francophone Studies* 12, 385–404.

Kennedy, S. and Trofimovich, P. (2013) First-and final-semester non-native students in an English-medium university: Judgments of their speech by university peers. *Language Learning in Higher Education* 3 (2), 283–303.

Kennedy, S., Foote, J.A. and Buss, L.K. (2015a) Second language speakers at university: Longitudinal development and rater behaviour. *TESOL Quarterly* 49 (1), 199–209.

Kennedy, S., Guénette, D., Murphy, J. and Allard, S. (2015b) Le rôle de la prononciation dans l'intercompréhension entre locuteurs de français lingua franca. *La Revue Canadienne des Langues Vivantes* 71 (1), 1–25.

Knoerr, H. and Weinberg, A. (2005) L'enseignement de la prononciation en français langue seconde: De la cassette au cédérom. *Revue Canadienne des Langues Vivantes* 61, 383–405.

Lappin-Fortin, K. and Rye, B.J. (2014) The use of pre-/posttest and self-assessment tools in a French pronunciation course. *Foreign Language Annals* 47 (2), 300–320.

Lindemann, S. (2002) Listening with an attitude: A model of native-speaker comprehension of non-native speakers in the United States. *Language in Society* 31, 419–441.

Lyche, C. (2010) Le français de référence: éléments de synthèse. In S. Detey, J. Durand, B. Laks and C. Lyche (eds) *Les Variétés du Français Parlé dans l'Espace Francophone: Ressources pour l'Enseignement* (pp. 143–165). Paris: Ophrys.

Marcoux, R. and Harton, M.E. (2012) *Et demain la francophonie: Essai de mesure démographique à l'horizon 2060*. Observatoire démographique et statistique de l'espace francophone. Québec: Université Laval.

Marcoux, R. and Konaté, M.K. (2011) Africa and the francophonie of tomorrow: An attempt to measure the population of the Francophonie from now to 2060. *African Population Studies* 25 (2), 215–225.

Munro, M.J. and Derwing, T.M. (1995a) Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning* 45 (1), 73–97.

Munro, M.J. and Derwing, T.M. (1995b) Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech* 38 (3), 289–306.

Organisation Internationale de la Francophonie (2010) *La Langue Française dans le Monde 2010*. Paris: Nathan Éditeurs.

Rossiter, M.J. (2009) Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review* 65, 395–412.

Segalowitz, N. (2010) *The Cognitive Bases of Second Language Fluency*. New York: Routledge.

Seidlhofer, B. (2013) *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.

Statistics Canada (n.d.(a)) Figure 1: Percentage of Canadian born (non-immigrants), foreign born (immigrants) and non-permanent residents in Montreal (CMA) [Graph] (last updated 15 October 2013). *NHS Focus on Geography Series*. Ottawa, Ont.: Statistics Canada. See https://www12.statcan.gc.ca/nhs-enm/2011/as-sa/fogs-spg/Pages/FOG.cfm?lang=E&level=3&GeoCode=462 (accessed 19 March 2016).

Statistics Canada (n.d.(b)) Detailed mother tongue, Montreal (CMA) [Table] (last updated 27 November 2015). *Census Profile*. See http://www12.statcan.ca/census-recensement/2011/dp-pd/prof/details/page.cfm?Lang=E&Geo1=CMA&Code1=462&Geo2=PR&Code2=24&Data=Count&SearchText=Montreal&SearchType=Begins&SearchPR=01&B1=Language&Custom=&TABID=1 (accessed 19 March 2016).

Storch, N. (2002) Patterns of interaction in ESL pair work. *Language Learning* 52 (1), 119–158.

Sturm, J. (2013) Liaison in L2 French: The effects of instruction. In. J. Levis and K. LeVelle (eds) *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference* (pp. 157–166). Ames, IA: Iowa State University.

Trottier, S. (2007) L'évaluation de la production orale chez les adultes en francisation: Analyse des critères de cinq enseignantes. Unpublished master's thesis, Université de Québec à Montréal.

VanWynsberghe, R. and Khan, S. (2007) Redefining case study. *International Journal of Qualitative Methods* 6 (2), 80–94.

Wells, C. (2013) Snap judgement: L'influence de l'origine ethnique, réelle ou imaginée, sur les évaluations des compétences en langue étrangère. *Linguistics*. See http://dumas.ccsd.cnrs.fr/dumas-00876398/document.

Wilkerson, M.E. (2013) The sound of German: Descriptions of accent by native and non-native listeners. *Die Unterrichtspraxis/Teaching German* 46 (1), 106–118.

Winke, P. and Gass, S. (2013) The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly* 47 (4), 762–789.

Zhang, Y. and Elder, C. (2014) Investigating native and non-native English-speaking teacher raters' judgements of oral proficiency in the College English Test–Spoken English Test (CET-SET). *Assessment in Education: Principles, Policy & Practice* 21 (3), 306–325.

Appendix: Empirical Codes, Examples and Frequencies of Coded Categories Used to Analyze Teacher-raters' Transcribed Verbatim Comments

Code	*Explanation*	*Example*	*Total frequency (%)*[a]			
			Acc	*Comp*	*Flu*	*CommEff*
AB	Raters' ability to evaluate construct	I find it hard to assess because I don't have enough to base my judgement on.	10	14	10	24
ACC	Linking construct to accent	I score her as more effective, but the existence of the accent will affect communication.	–	6	0	0
ATT	Attitude towards/familiarity with language or accent	Typically, the Chinese accent doesn't bother me, I'm so used to it.	5	1	0	0
CO	Oral comprehension (speaker)	He's not following at all what she says.	0	6	2	22
COMP	Linking construct to comprehensibility	Communicative effectiveness goes quite a bit with comprehensibility.	1	–	0	0
CONT	Continuity (smoothness) of speech	There are almost no hesitations.	1	2	17	3
CP	Participation in interaction	He is very passive, she leads all the way.	3	5	8	40
CS	Communication strategy use	She validates the information.	2	6	4	58
CTX	Role of context	If I had to judge with respect to their readiness for the workplace, I would judge differently.	0	1	3	5
DEB	Speech rate	The conversation moves right along; they move forward quite fast.	0	0	6	0
DIS	Discourse ability	She describes it well.	0	5	3	4
EMP	Speaker's empathy	She wants to reassure her partner and tells her: 'You had the hard part …'.	0	0	0	0

EX –	Exclusion of element for rating	She has a strong accent but it doesn't affect her comprehensibility.	0	6	0	5
FIX[b] –	Target for future improvement	She should work on pronunciation.	1	0	0	5
FOS	Fossilization	It will be very hard for her to change that at this stage, even with practice.	2	0	0	0
GLOB	Global evaluation of speech	They're really good!	5	17	19	24
L1REF	Native speaker norm	I'm trying to figure whether a Francophone would understand.	0	24	3	3
LF	Lingua franca	They understand each other in spite of their mistakes.	0	15	3	7
LX	Lexis	He should work on his vocabulary.	0	13	2	13
LX – A	Lexical access	She's searching for her words.	0	3	11	2
LX – K	Lexical knowledge	She lacks all functional words.	1	3	2	4
NAT	Nature of speaker's L1	There are no articles in Russian and so she omits them systematically in French.	17	6	1	1
NIV[b]	Proficiency level of speaker	They are at a very basic level.	4	1	2	4
NIVC	Comparison/contrast with other speaker	She communicates better than him.	11	13	13	22
P	Phonological elements	In both cases, their pronunciation is awful.	4	23	0	7
PS	Segmental	She substitutes [t] for [d].	12	13	1	2
PSS	Suprasegmental	She has a flat intonation.	5	2	3	0
SA	Salience of accent	Her accent is very marked.	19	2	0	0
SG	Syntax and morphology	She omits conjugations.	6	12	3	4
TASK	Role of task	My assessment would have been different if this had been a conversation instead of a goal-oriented task.	5	7	8	17
Total			104	188	103	243

Notes: [a]Acc = accentedness, Comp = comprehensibility, Flu = fluency, CommEff = communicative effectiveness. [b]GLOB, NIV and FIX not included.

13 Pronunciation Assessment in Asia's World City: Implications of a Lingua Franca Approach in Hong Kong

Andrew Sewell

Introduction

Sitting in a tapas bar in Hong Kong, I overheard a conversation about English accents in contemporary social life: 'Look like one thing, sound like another, and live somewhere else', said a woman to a British-sounding but Asian-looking young man. Her observation captures rather neatly the diversity and unpredictability of language use in the age of globalization, which is also the age of 'super-diversity' (Vertovec, 2007). The audible and visible phenomena of superdiversity arise from global flows of various kinds: people, goods, capital and information (Held *et al.*, 1999). The forms and uses of English, the de facto language of globalization, are also affected by these flows. The transformations and recombinations involved are highly complex, however. It is not merely a matter of there being 'local' Englishes, as is sometimes suggested by a World Englishes or a conventional diversity perspective; the local and the global are now intimately connected. A superdiversity viewpoint also acknowledges the important role of new media and new technologies of communication (Blommaert & Rampton, 2011: 3).

The pronunciation of English is also subject to the effects of superdiversity. It is of course affected by people's other languages, but we can also hear the effects of the globalization of accent features and identity orientations. As one example, it no longer comes as a surprise to encounter students who have acquired English accent features from the media (e.g. see Zhang, 2003, in the case of American accents in China). Under these conditions, the challenge for

pronunciation teaching and assessment lies in navigating the local/global polarity and making pedagogical sense of the complexity. One solution is provided by a lingua franca approach, and this chapter outlines one interpretation of such an approach. It first surveys the findings of research into intelligibility in lingua franca contexts. It then links the findings with the theoretical construct of functional load, which is elaborated and framed within functionalist approaches to language and communication. Some pedagogical implications are then identified, using Hong Kong as a case study. The lingua franca approach is shown to have specific indications for the prioritization of features in the pronunciation descriptors for a local test of English. It also has more general implications for the teaching and assessment of pronunciation in a globalized world.

Navigating the local and the global in pronunciation teaching and assessment

Derwing (2008: 348) rightly points out that understanding the 'milieu' in which students find themselves is 'critical in designing a curriculum that adequately addresses pronunciation needs'. In the 21st century, understanding that the milieu contains both local and global elements is at the heart of the matter. Designing curricula also means adopting an interdisciplinary approach and taking account of both linguistic and sociolinguistic factors. Flexibility and adaptability are essential requirements for effective communication in a globalized world, and in many contexts it is no longer useful to insist on the reproduction of a native speaker language system (or any other pre-formed system, for that matter). In assessment, what matters is the ability to successfully perform 'linguistically mediated tasks' (Hall, 2013: 227), ones in which the construct of interest is 'the performance of the task itself' (Long & Norris, 2009: 139).

The increasing use of pair work in language assessment practices (see Elder & Harding, 2008) is one way in which these tasks are being introduced. Examples might include 'information gap' tasks that require the sharing of information, or other tasks involving communication and cooperation (see Nunan, 2004: 174). But the theorization of language and communication under conditions of diversity and unpredictability is still in its infancy. Sussex and Kirkpatrick (2012: 224–225) posit the existence of two poles of language use, those of system-entity-edifice (SEE; representing the more systematic, predictable contexts or aspects of language use) and lingua franca English (LFE; representing the emergent, less predictable contexts or aspects). The paradigm case of SEE is formal written language, while LFE is typified by informal spoken language, and by situations in which interlocutors come from different language backgrounds. Canagarajah (2007: 933) characterizes the revised view of 'competence' associated with an LFE perspective: it is not so much 'applying mental rules to situations', but rather 'aligning one's resources with situational demands'.

Sussex and Kirkpatrick (2012: 225) note that the 'extent and way in which the system and emergent frameworks can co-exist and collaborate represent a major challenge for research'. SEE and LFE do not exist in an either/or relationship, and many spoken tasks and interactions will still involve fairly predictable language forms. As Sussex and Kirkpatrick (2012: 225) acknowledge, what speakers bring to each new situation is not 'a linguistic tabula rasa'. To put it another way, although interaction involves emergent language forms and unpredictability, much of the variation takes place against a backdrop of intelligibility, as noted by Parkin (2012: 74): 'speakers juggle the limits of face-to-face intelligibility at any one time with new styles of expression made up of ever changing linguistic resources.'

Navigating the currents of global English therefore means several things. It means accommodating diversity without succumbing to simplistic notions of 'local English'. It means prioritizing adaptability and flexibility over the ability to reproduce a predefined system, while at the same time acknowledging both the systematic and the emergent qualities of interaction and recognizing the continuing importance of intelligibility (see Munro, 2013). A promising solution is provided by a 'lingua franca' approach to pronunciation teaching and assessment.

Lingua franca approaches to teaching and testing

A lingua franca approach focuses on transnational interactions, and adopts intelligibility (rather than nativeness) as a key orienting principle. In lingua franca research, intelligibility tends to be investigated using corpora of natural conversations among nonnative speakers. Instances of misunderstanding are identified via observer reports of communication breakdowns (e.g. Jenkins, 2000) or participant reports of misunderstanding (e.g. Deterding, 2011, 2013). The number of pronunciation-related instances of misunderstanding in these corpora has ranged from 27 (Jenkins, 2000) to 158 (Deterding, 2013). These instances are then analyzed in terms of their possible causes. A landmark text in the development of the corpus-based, lingua franca approach to intelligibility was Jenkins (2000), one of the first to study intelligibility in interactions between nonnative speakers of English. Jenkins's research identified the well-known lingua franca core (LFC) of pronunciation features that contributed to intelligibility in these interactions. Essentially, these were most consonantal features and contrasts (with the exception of the dental fricatives), vowel contrasts maintained by length, and nuclear stress. Other features, notably many suprasegmental features, were classified as 'non-core'. This means that they are less essential for the achievement of intelligibility.

Although ground-breaking in many ways, the LFC was never intended to be a definitive statement of the factors affecting intelligibility. There has been considerable discussion and criticism (e.g. Dauer, 2005; see also the

edited volume by Dziubalska-Kolaczyk & Przedlacka, 2005), and I will not present another detailed evaluation here. Instead, I would like to propose some possible explanatory factors, and to see whether these are relevant to the findings of more recent research studies. The main argument I make is that it may be profitable to view the findings through the lens of *functional load*. This is something of a problematic concept (see Brown, 1991; Surendran & Niyogi, 2006), and in this chapter I make a distinction between narrow and broad senses of the term in order to clarify and elaborate its meaning.

In its narrow sense, functional load refers to the amount of work done by different phonemic contrasts.[1] It can be measured and compared by combining various criteria, most notably the number of minimal pairs that a phonemic contrast serves to differentiate. For example, the functional load of the /ʊ–uː/ vowel contrast is lower than that of the /ɪ–iː/ vowel contrast, as there are fewer minimal pairs in the former case (*full* and *fool* being one of the few confusable pairs). Calculations usually result in broadly similar rankings (see Brown, 1991; Catford, 1987). This version of functional load suggests a possible reason why the dental fricatives /ð/ and /θ/ are not part of the LFC – there are relatively few minimal pairs involving these sounds, and the probability that substitutions will reduce intelligibility tends to be lower.

The concept of functional load can be extended to include a broader sense, that of the informational relevance of linguistic features or classes of such features (see Surendran & Niyogi, 2006). From the viewpoint of functionalist approaches to language (e.g. Bybee, 2001; Givón, 1995), every linguistic device has a function (Bardovi-Harlig, 2007). The relative importance of different devices can be compared:

> if an adverb such as *yesterday* is the only indicator in a sentence that an event happened in the past, then the functional load of the adverb is high. If the sentence also employs past-tense verb morphology to indicate the time frame, the functional load of both the adverb and the verbal morphology is less than either one occurring alone. (Bardovi-Harlig, 2007: 59)

Whether it is applied to lexico-grammar or phonology, the broad version of functional load therefore indicates some of the constraints on the reduction or alteration of information, and suggests where problems are likely to occur. Although Surendran and Niyogi (2006) take a computational approach, precise measurement is not essential in order to apply the concept and obtain insights into the findings of lingua franca intelligibility research. For example, a consistent finding of such research is that consonants are more important than vowels for intelligibility. Deterding (2011: 93–94) lists 11 instances of intelligibility problems occurring in conversational interactions between speakers from the Association of Southeast Asian Nations (ASEAN) region. Nine involve the substitution or deletion of consonants

(apart from the dental fricatives); an example is the word *nearby* being pronounced with initial [l] by a Cantonese speaker from Hong Kong. The overall conclusion of Deterding's (2013: 91) corpus-based study was that 'the greatest impact on intelligibility comes from consonants, which is consistent with the LFC proposals'.

The interactions between consonants and vowels in processes of word recognition are extremely complex, but it is possible to suggest that consonants generally have a greater functional load and are more important for intelligibility. Cruttenden (2014) draws on functional considerations and concludes that vowel contrasts are 'less crucial' to intelligibility in English than consonant contrasts. It is important to bear in mind, however, that statements such as these are not necessarily concerned with the properties of an abstract language *system*, as in the narrow or traditional sense of functional load. Instead, corpus-based lingua franca studies are more concerned with language *practices* (i.e. how actual speakers and listeners rely on certain features in the moment-to-moment achievement of intelligibility). It may be worthwhile to elaborate the concept still further and say that consonants appear to have a greater *effective* functional load in lingua franca communication, regardless of their theoretical properties. There is at least some statistical evidence for the greater functional load of consonants in English. Cutler (2005), for example, computed that for words of various lengths, there are about 2.2 times as many lexical neighbours if a consonant is replaced (*cat* becoming *mat*) than if a vowel is replaced (*cat* becoming *cot*).

The broad sense of functional load also helps to explain another finding of the LFC and similar studies, namely that initial consonant clusters do more 'work' in terms of intelligibility than final ones. Deterding (2013: 90) notes that the loss of the second consonant from initial clusters such as [pl] and [fr] was among the 'biggest problems' with consonants, while the omission of [t] or [d] from final clusters was relatively unproblematic. Psycholinguistic research suggests that word recognition proceeds in temporally linear fashion from the beginning of the word (Marslen-Wilson & Zwitserlood, 1989). Modifications to clusters at the beginning of words therefore seem more likely to cause problems than those situated elsewhere. In his 'information theory' approach to redundancy in English, Shannon (1951: 55) makes the point well: the beginning of words is where 'the line of thought has more possibility of branching out'. Functional factors in information processing also help to explain why Jenkins (2000: 142) identifies epenthesis as being preferable to deletion in clusters (e.g. as an intelligible pronunciation of *black*, [bəlæk] would be preferred to [bæk]). It is generally better to add information, perhaps allowing the listener to extract what is needed, than to remove it (see Lin, 2003).

One might therefore wonder why functional explanations seem to be resisted in lingua franca intelligibility studies. It is true that there is little direct evidence of the relationship between functional load and intelligibility,

although Munro and Derwing (2006) concluded that their study offered 'preliminary confirmation' of the functional load hypothesis (namely, 'errors' involving contrasts with a high functional load are more likely to reduce intelligibility). A more probable reason for the neglect of functional explanations is that research in the English as a Lingua Franca (ELF) research paradigm (e.g. Jenkins, 2000; Seidlhofer, 2004) is not only concerned with the linguistic aspects of intelligibility; it also has a sociolinguistic, activist orientation in terms of wishing to change perceptions of the nonnative speaker. From this viewpoint, the problem with functional load is that it suggests reliance on an existing system. This is perhaps seen to have a centripetal, native speaker influence that conflicts with the centrifugal, de-centring aims of ELF research but, as noted above, the broader sense of functional load is not incompatible with these more practice based orientations.

The case for changing perceptions of 'error', and of so-called nonnative speakers in general, is persuasive (see Lindemann, this volume). Under conditions of globalization and diversity it is difficult to justify teaching or assessment approaches that treat every departure from a monolithic standard as an error. However, it has to be pointed out that the rhetorical construction of ELF as a distinct entity has led to a certain over-polarization of the issues. There is a tendency to exaggerate the differences between native and nonnative speakers, and overstate the extent to which the lingua franca approach involves different targets. Yazan (2015: 203), for example, asserts that the LFC 'repudiates adherence to native-speaker norms'. If one thinks in terms of phonological contrasts, this is a puzzling statement. According to the LFC research, the vast majority of the consonantal contrasts in 'native speaker' models such as Received Pronunciation (RP) and General American (GA) are in fact also needed for international intelligibility. The LFC arguably represents a conservative orientation towards pronunciation teaching.

Naturally, maintaining consonantal contrasts does not mean that they have to be pronounced in the same way, only in ways that allow listeners to make distinctions that are conducive to intelligibility. There is of course a huge amount of inter-speaker accent variation, as well as intra-speaker variation in terms of speech styles and speech rates. A lingua franca approach aims to accommodate this variation. But it does not suggest that 'anything goes', and uses intelligibility as a central criterion in distinguishing between problematic and unproblematic accent features. In terms of its application to pedagogy, the lingua franca approach can therefore be seen as a continuation and refinement of the intelligibility principle (see Levis, 2005). This principle has a long history in pronunciation teaching, dating back at least as far as Abercrombie's (1949) 'comfortable intelligibility'. The lingua franca approach is not particularly new (see Munro & Derwing, 2015), and in terms of linguistic targets it is not particularly different.

In another case of over-polarization, Cook (2011: 149) draws on ELF research and concludes that 'the phonology of ELF is different to that of native

English'. This is certainly true in a trivial sense; the phonology of *any* group of speakers must be different from that of any other group, once we move beyond the outdated essentialism of believing that groups have inherent characteristics or share a 'common underlying system' (e.g. Hung, 2000, in the case of Hong Kong English). Reflecting on diversity, we realize that generalizations about 'native' or 'nonnative' speakers are difficult to sustain in today's world. 'Native speaker' contexts are themselves extremely diverse as a result of migration and inequality, among other factors. To a large extent we are all lingua franca users, even if we live in so-called monolingual environments. A consequence of diversity (and of sophisticated ways of seeing diversity) is the realization that 'every language is a multiplicity of codes' (Croft, 2000: 92).

A lingua franca approach provides some navigational aids in the midst of this complexity. If we accept that functional factors help to explain many of the findings of lingua franca intelligibility studies, then what these studies do is to identify the features that all speakers, regardless of background, tend to rely on to maintain intelligibility. If this sounds like an unacceptably centripetal statement in the context of what has been said about diversity, there are two qualifications. First, there is a great deal of accent variation taking place, as noted above. Secondly, if we reflect on what it is that creates the apparently gravitational force of the core, there is no need to see it as representing lingering 'native speaker' influences. It is instead possible that written language and worldwide literacy operate as centripetal forces on pronunciation, especially in more formal contexts of use. While accepting diversity and emergent patterns, shared knowledge of written forms is one of the 'anchoring practices' (Swidler, 2001) that affect and constrain linguistic variation in international communication. There has been a recent upsurge of interest in the question of how knowledge of orthography affects perception, production and acquisition. In their overview, Bassetti *et al.* (2015) note that assumptions of the 'primacy of speech' and of the separateness of spoken and written language have long dominated both research and teaching. This is changing in favour of the realization that speech and writing are not separate, but represent 'closely related, and often complementary, systems' (Katamba, 2005: 221).

Pronunciation Assessment in Hong Kong: Room for Improvement?

If the findings of lingua franca research into intelligibility are linked to functional factors, they gain added explanatory power and are therefore more pertinent to discussions of norms in teaching and assessment. But to repeat what was said above about over-polarization in ELF research, there is nothing new in applying the intelligibility principle in this way. What the lingua franca approach does perhaps offer, in terms of novelty, is that the focused

data from intelligibility studies allow for a more detailed, feature-based evaluation of pronunciation teaching syllabi and assessment descriptors. Combined with the findings of more experimentally-controlled studies, it may therefore suggest changes in terms of priorities and overall orientations.

In this section I will determine the scope for incorporating a lingua franca approach into the assessment of pronunciation in Hong Kong. There are several reasons why Hong Kong makes an interesting lingua franca case study. As around 95% of the population speaks Cantonese, English plays a fairly limited role in internal communication (but see Evans, 2011, for an alternative view). What matters in Hong Kong – 'Asia's World City', as it is currently branded – is people's ability to communicate with a range of interlocutors from different linguacultural backgrounds. Although recent studies have evaluated local pronunciation teaching and testing materials through the lens of an ELF approach (e.g. Chan, 2014), these have tended to take the overly polarized positions noted above, and have neglected the possible functional explanations for lingua franca patterns of communication.

It is necessary to begin by trying to characterize the current situation – not an easy task, when descriptors and other curriculum documents may not reflect what is actually going on. The history of the pronunciation descriptors in a local English examination called the Language Proficiency Assessment for Teachers of English (LPATE) provides a useful starting point. The test was developed in order to benchmark the proficiency level of Hong Kong's English teachers (see Coniam, 2013). All prospective teachers have to pass the test in order to teach in primary or secondary school classrooms. The test includes five papers (Reading, Writing, Listening, Speaking, and a Classroom Language Assessment or CLA), and pronunciation is assessed in the last two of these; in the LPATE as a whole, Band 3 is the minimum level that candidates need to attain. Early versions of the LPATE pronunciation descriptors were oriented towards nativeness (or accentedness), rather than intelligibility. Pronunciation at the 'above the benchmark' level was characterized as being 'completely error-free with no noticeable L1 characteristics' (Coniam & Falvey, 2002: 23). Current versions, however, appear to be more compatible with a lingua franca approach (see Table 13.1).

Although the 'nativeness' orientation of the descriptors has been abandoned, from a lingua franca perspective the problematic terms here are 'errors' and 'natural' – what constitutes an 'error', and who decides what 'natural' means? (See also Harding, this volume.) After analyzing LPATE examiners' assessment reports (see Sewell, 2013), I concluded that the identification of 'errors' does in fact reflect a general orientation towards intelligibility rather than accentedness. This was clearly visible from the analysis of examiner comments relating to segmental features, where the levels of agreement with the LFC criteria in the areas of vowels, consonants and consonant clusters were 81%, 99% and 97%, respectively. The relative absence of comments relating to the dental fricatives /ð/and /θ/ illustrates the intelligibility orientation of the

Table 13.1 Descriptors for pronunciation, stress and intonation in the LPATE examination handbook

LPATE band	*Descriptor*
5	Reads in a fully comprehensible way with no systematic errors in pronunciation and uses stress and intonation in a very natural way.
4	Reads in a comprehensible way with few systematic errors in pronunciation and uses stress and intonation in a mostly natural way.
3	Reads in a generally comprehensible way, although may make errors in pronunciation. Uses stress and intonation to convey meaning, although may occasionally sound unnatural.
2	Does not read in a consistently comprehensible way due to errors in pronunciation, stress and intonation and speech is frequently hesitant.
1	Makes frequent errors in pronunciation, stress and intonation which cause confusion for the listener.

Source: Adapted from HKEAA (2011: 71).

LPATE examiners. The LFC suggests that substitutions of these sounds are unproblematic, and although they are prevalent in the spoken English of Hong Kong students, dental fricative substitutions were hardly ever mentioned in the assessment reports. In this area of assessment at least, one gets the impression that 'communicative effectiveness' has already replaced 'rigid adherence to SE [Standard English] norms' (Elder & Harding, 2008: 3–4).[2]

Turning to suprasegmental features, the general indication of the LFC is that suprasegmentals are 'non-core' features, with the exception of nuclear stress and the possible exception of word stress. Jenkins (2000: 150) concedes that word stress is something of a grey area. Applying a functional perspective may be useful: other things being equal, stress modifications that change vowel quality are more likely to cause problems, as are modifications in bisyllabic words, compared with tri- or multi-syllabic words. Thus pronouncing the word *written* with primary stress on the second syllable may make it sound like *retain* to some listeners, especially if there are associated vowel effects. Field (2003) believes that stressed syllables may serve as 'islands of reliability' for listeners, perhaps because they serve as initial cues for processes of word recognition (Grosjean & Gee, 1987). In terms of intelligibility, the suprasegmental level of word stress is thus related to the segmental level (Kohler, 2011).

But while this may suggest that word stress is actually a core feature in terms of maintaining intelligibility, this should not result in the naturalization of all aspects of native speaker patterns. For example, the LPATE descriptors for suprasegmentals state that a top-scoring candidate 'uses stress and intonation in a very natural way'. The problematic term here is 'natural'. It is highly subjective, and there is no connection between naturalness and intelligibility. If we look at examiners' comments in more detail, they

sometimes refer to speech phenomena that are in fact unlikely to affect intelligibility. To consider word stress first of all, the comments included 'giving stress to the weak vowel sounds as in "chocolate", "carrot" and "ceremony"' (HKEAA, 2005: 14) and 'wrong syllable stress as in multi-syllabic words like "informative" and "superlative"' (HKEAA, 2009: 13).

In the first case, assuming that 'giving stress' to the weak vowel sounds of *chocolate* involves retaining its overall stress pattern, a possible pronunciation in Hong Kong might be ['tʃɒkoʊleɪt]. There are three syllables, as opposed to two in most 'dictionary' representations (e.g. /'tʃɒklət/). But if primary stress is placed on the first syllable, such a pronunciation seems unlikely to reduce intelligibility. It may even be argued to *increase* intelligibility in international communication (see Deterding, 2010), partly because this kind of 'spelling pronunciation' makes the spoken form more like the written form. If we consider what happens in momentary instances of unintelligibility, we can appreciate that listeners may engage in a process of reconstruction, trying to work out what the speaker said. Recalling what was noted above about the effects of literacy, it is also likely that listeners will visualize possible spellings of the word, assuming that they have time and inclination to do so (see Cutler *et al.*, 2010; Ong, 2002). Using full vowels in *chocolate* facilitates the drawing of sound-spelling analogies with other words, such as *late.* Generally, ['tʃɒkoʊleɪt] seems more likely to be intelligible, especially if the word is unfamiliar for the audience in question.

In the second case, in Hong Kong one can often hear multi-syllabic words being given 'non-standard' patterns of stress such as *inforMAtive* or *communiCAtive*. Although the stress pattern may be unfamiliar to some listeners, the multi-syllabicity of the words means that it probably does not matter; multisyllabic words have fewer competitors in their lexical neighbourhood, to use the term of Luce and Pisoni (1998). Again, from a functional, lingua franca perspective, neither the amount nor the kind of phonological information is likely to cause intelligibility problems. There are few or no grounds for penalizing such pronunciations, or for trying to 'naturalize' pronunciation around native speaker norms.

Similar arguments apply to intonation. The LFC includes nuclear stress as a core feature, suggesting that other aspects of intonation may not contribute much to intelligibility – including those which might sound 'natural' from a nativeness orientation. For example, the LPATE examiner comments indicate that assessors did not look favourably on candidates breaking a 'rule' of intonation, namely that yes/no questions have rising intonation, while wh-questions generally fall: '[s]ome candidates tended to use the rising tone for all question types, as in "How do you spell it?"' (HKEAA, 2006: 11). Analysis of actual conversation shows that native speakers often break this rule, and in any case it seems unlikely to affect intelligibility.

Another suprasegmental area noted in the examiner comments is that of linking: 'there seemed to be a general lack of attention to the linking of

sounds as in "think about it"' (HKEAA, 2008: 12). Although the complete absence of linking would tend to make speech rates extremely slow, 'non-linking' can be defended on the grounds that it makes word boundaries clearer and enhances intelligibility. Native speakers also vary the amount of linking for rhetorical or informational effects.

Considering the LPATE examiner comments as a whole, one of the indications of a lingua franca, intelligibility-based approach is that many supra-segmental features can be given a lower priority. As mentioned above, at the very least it can be argued that candidates who use 'non-standard' features such as word stress modifications in multi-syllabic words, or alternative intonation patterns, or who 'fail' to use connected speech phenomena such as linking, should not be penalized. Taking a global perspective, research suggests that syllable-timed (as opposed to stress-timed) rhythm is a characteristic of many so-called 'new Englishes'. For example, Setter (2006) found that Hong Kong speakers generally showed less difference than British speakers in the relative duration of stressed and unstressed syllables.

The pronunciation descriptors of other examinations in Hong Kong also indicate that there is scope for a lingua franca approach. The Hong Kong Diploma of Secondary Education (HKDSE) was introduced in 2012 as part of a new curriculum. The published descriptors for pronunciation are not very detailed, but they suggest a general orientation towards nativeness or an unanalyzed 'standard', rather than intelligibility. To attain Level 6 or 7 on the seven-point scale, candidates must 'pronounce all sounds/sound clusters and words clearly and accurately' (HKEAA, 2013: 165). We cannot be sure how these descriptors are interpreted, but they appear to assume that pronunciation merely reproduces written forms. Pronouncing *all* sounds, for example in final consonant clusters, is something that no speakers of English do (Schreier, 2009). A lingua franca approach would problematize the nature of 'accurate' pronunciation here, acknowledging the possibilities for variation and drawing on evidence from intelligibility studies.

It must be noted, however, that adopting a lingua franca approach does not mean going to the opposite extreme by reifying either a 'lingua franca model' or a 'local model'. A further indication of the lingua franca approach is that the very concept of a 'model' is too limiting. Elder and Harding (2008: 4) point out that, rather than invoking 'English as an International Language' or 'Standard English' as constructs, they are instead 'arguing for a contextualized description of what we are attempting to measure'. Intelligibility suggests that *features* – rather than models or varieties – should form the primary units of analysis in teaching and assessment.

Once again, this kind of nuance tends to evaporate in the over-polarized climate of much research in the ELF paradigm. For example, in surveying the overall situation in Hong Kong, Chan (2014) claims that local assessment guidelines and textbooks are oriented towards native speaker norms, and argues that 'a key step in an ELF approach is a codification process which

targets the pronunciation features of the educated HKE [Hong Kong English] speakers with reference to the LFC' (Chan, 2014: 149). Over-polarization results in the construction of an idealized and homogeneous 'Hong Kong English', which leads Chan to downplay the indications of the LFC. He states that the pronunciation targets of a local textbook refer 'exclusively' to 'NS-correctness' (Chan, 2014: 161). But of 13 segmental features taken from the book and listed in a table (Chan, 2014: 162), a large majority turn out to be problematic for intelligibility, according to the LFC criteria. For example, initial cluster simplification in words like *clothing* is one of the features mentioned (a possible Hong Kong pronunciation would be ['kouðɪŋ]). However, both empirical research (e.g. Deterding, 2011, 2013; Jenkins, 2000) and functional considerations suggest that this kind of simplification is indeed likely to be problematic. Highlighting this and other features in teaching and assessment is not being 'NS-centred'; on the contrary, it closely reflects the findings of lingua franca intelligibility studies.

In evaluating the possible contributions of a lingua franca approach, then, we should be wary of the tendency to see native/nonnative, lingua franca/non-lingua franca dualisms when the actual situation is more complex. Chan's own interpretation of a lingua franca approach involves accepting local accent features, but this definition of 'local' draws uncritically upon the lists of 'typical' features found in descriptive studies such as Hung (2000). This veers too far towards the conventionally local. Far from de-centring English, it merely replaces the irrelevant centripetality of the native speaker model with the limiting centripetality of an undifferentiated local model. Many of the 'typical' or 'distinctive' phonological features identified by linguists in the World Englishes paradigm turn out to have a rather limited distribution within the population. In the corpus of Hong Kong media English collected by Sewell and Chan (2010), some of the features listed by Chan (2014), such as the substitution of /v/ with [w] and the 'conflation' of [n] and [l], were used by less than 20% of the speakers. A possible, functional explanation is that experienced speakers learn to avoid features that reduce intelligibility. Additional research by Sewell (2015) showed that /v/-substitution reduced intelligibility even in local, intra-ethnic communication. Although 'function' is not the only explanation, the functionalist viewpoint maintains that 'meaning-making efforts on the part of the learner are a driving force in ongoing second language development' (Mitchell *et al.*, 2013: 188).

Implications of a Lingua Franca Approach

In discussing possible implications, it should of course be emphasized that further research and theorization are needed in order to substantiate the findings of lingua franca studies. Nevertheless, the implications of a lingua franca approach to pronunciation assessment in Hong Kong are of two main

kinds. First, there are specific, features-based indications, which mainly apply to pronunciation scale descriptors. Secondly, there are more general implications for the overall approach to teaching and assessment.

In terms of features, a lingua franca approach suggests that some of the 'native speaker' features mentioned in descriptors (such as those of the LPATE) can probably be given a lower priority – assuming that they are currently prioritized, which seems doubtful in some cases. The LPATE examiners hardly ever mentioned substitutions of the dental fricatives, for example, perhaps because they were less noticeable or did not interfere with actual intelligibility (it is also possible that substitutions were so ubiquitous as to make commenting on them pointless).

To some extent, one could simply rely on this fact to accommodate local variation: examiners are automatically relying on intelligibility as a guide, because they bring their accumulated experience of English communication with them. In the LPATE examiner reports, the vast majority of comments refer to features that would be expected to reduce intelligibility, according to the LFC criteria (these included substitutions and deletions of consonants, apart from the dental fricatives, and modifications of initial consonant clusters). The case of the dental fricatives suggests that non-standard segmental features that do not reduce intelligibility will tend to 'fly under the radar' of examiners' attention.

On the other hand, the LPATE examiner comments indicate that other non-standard features may be noticeable, and penalized, despite their being unproblematic for intelligibility (the absence of vowel reduction is a case in point). Here the lingua franca approach runs into a possible problem, in that sociolinguistic factors such as stigmatization may tend to override the 'rational' criterion of intelligibility.

My approach to the resolution of this problem is a pragmatic one. It aims to achieve a certain amount of change in terms of local orientations towards teaching and testing pronunciation, while not antagonizing local language users or gate-keeping institutions. Attitude surveys and a consideration of the language-ideological landscape suggest that conservative views predominate, and it would be unwise to ignore the views of stakeholders – students, parents, teachers and administrators, in addition to the wider community – by uncritically accepting so-called 'local' features, as Chan (2014) appears to do. Instead, the strategy that I adopt here takes account of both the linguistic dimension of intelligibility and the sociolinguistic dimension of acceptability. It negotiates acceptability partly by considering the noticeability of features; it is argued that those that escape most people's awareness might as well be accepted. In addition to intelligibility and noticeability, there is also the question of learnability. If features are difficult to acquire, as the dental fricatives are for many learners, the case for acceptance is further strengthened.

What enters or escapes people's awareness, and what is easy or difficult to learn, depends on local conditions. Proposals for local examinations may not

always apply to international ones, although there are some principles with more general applicability. In the following paragraphs I will identify and discuss some of the possible candidates for acceptance in Hong Kong. Acceptance may mean that certain 'non-standard' features are currently being penalized in assessment, and need not be. Or it may mean that these features are not currently remarked upon, but are still worth identifying as being unproblematic.

In terms of segmental features, neither of the dental fricatives appear to be worth prioritizing (in other words, substitutions can be accepted). They do not have high functional loads, and substitutions have not often been noted as sources of intelligibility problems. If we consider the noticeability of the two sounds, there is a slight difference between them: substitutions of the voiceless dental fricative /θ/ are perhaps more likely to be noticed, because possible contexts for the sound include stressed syllables (such as the high-frequency content words *three* and *think*). It might therefore be given a slightly higher priority in teaching. Contexts for the voiced dental fricative /ð/ occur more frequently (as in *the* and *them*), but several factors combine to make substitutions less problematic: the low functional load of these so-called 'function' words; the fact that they are usually unstressed and less noticeable; and the possibility that the ubiquity of substitutions might have a desensitizing effect on the listener. To the criteria of intelligibility and noticeability, learnability can be added in this case: dental fricative substitutions represent adaptations of sounds that pose particular problems for many learners of English.

The treatment of final consonant clusters in descriptors may be another area for review. The LFC findings and the functional approach to intelligibility indicate that simplifications of final clusters are less likely to reduce intelligibility than simplifications of initial or medial clusters. Although this was acknowledged in the overall pattern of examiner comments, and although examiners will tend not to notice many of the simplifications that occur, the HKDSE descriptors appear to be too strict. The LPATE descriptors are less rigid, but the examiners are probably right to draw attention to final cluster simplification in monosyllabic words such as *paint* (pronounced as *pain*; HKEAA, 2003). But functional considerations suggest that the more syllables there are, the less likely it is that such modifications will cause problems. This is particularly so when simplification occurs in unstressed syllables. For instance, it is common for speakers in Hong Kong to delete /t/ or /d/ in the final clusters of words like *department* or *commitment*. This is unlikely to affect intelligibility and probably passes unnoticed much of the time. I doubt that many examiners would notice this or attempt to penalize it, but it is perhaps worth noting as an 'under the radar' feature of local pronunciation that can be accepted. As a practical recommendation, scale descriptors could attempt to focus on problematic simplification (i.e. in monosyllabic words), rather than giving the impression that all forms of simplification constitute 'errors'.

Turning to suprasegmental features, the main implication of a lingua franca approach is that these are generally less important in terms of their contribution

to intelligibility. As noted above in the case of *chocolate*, the tendency not to reduce vowels in unstressed syllables is a strong candidate for acceptance – the absence of reduction does not reduce intelligibility, and may actually increase it in international communication. In addition to the theoretical arguments advanced above, there is some empirical support for this view; Field (2005) found that 'full-quality' syllables assisted word recognition for both native and non-native listeners, and links this to the closer relationship with the orthographic form. In much the same way, word stress does not always need to follow the dictionary pattern. The low functional load of individual syllables in multi-syllabic words such as *informative* means that changes to the 'normal' pattern are unlikely to affect intelligibility. One could speculate as to whether *infor-MAtive* is more logical and regularized, and therefore likely to become dominant in future. Pronunciation descriptors and examiner guidelines could also orient themselves towards intelligibility, rather than nativeness, in this area.

From a lingua franca perspective, there is no obvious reason to insist on native speaker intonation patterns in questions. What seems to be more important is the listener's ability to detect intonation features with a high functional load, such as nuclear stress (as indicated by the LFC) and perhaps contrastive stress in general. Here intonation overlaps with other suprasegmental areas such as sentence stress, and more research is needed in order to assess the actual functional value of these features in different contexts of communication.

If we accept Elder and Davies's contention that nonnative accents may need to be incorporated in language tests (2006), a lingua franca approach centred on intelligibility provides a way of evaluating these accents, as well as native speaker accents. Field (2004) observes that a range of different 'standard' accents from around the world is more appropriate than the uncritical adoption of 'local' accents; the application of a lingua franca approach allows the standard to be inflected with local, intelligible variants. This is in fact what examination boards seem to be doing. In Hong Kong, Chan (2014) describes the local accents used in the HKDSE as being predominantly 'RP' on the grounds that they do not use 'typical' local features, but this again reveals the binary, either/or nature of a certain type of 'lingua franca' position. As most of the LPATE examiner comments suggest, the examination board seems to be balancing the need for a local perspective with the need for intelligibility. Whether this will actually promote international intelligibility is a complex question, because additional factors such as accent familiarity also need to be considered (see Ockey & French, 2014). It is possible that lingua franca intelligibility research mainly suggests the nature of the 'minimum threshold level' identified as being important by Rajadurai (2007: 102). Beyond this threshold, other speaker and listener variables will still have effects on intelligibility.

Moving from features to strategies, and broadening the discussion from testing to pronunciation pedagogy in general, it is possible that a more significant contribution of the lingua franca approach lies in its overall orientation and philosophy. A lingua franca approach has the goal of 'repertoire

expansion' rather than 'error eradication' (Ferguson, 2009: 130). I take this to mean that there is a more tolerant, and yet critical, attitude towards variation in general. When they are ready, students are encouraged to explore the sociolinguistic and pragmatic significance of different forms, rather than treating all departures from a notional standard as errors. The standard is not ignored, and in fact learners are empowered by becoming more aware of the standard/non-standard polarity and its significance (Andrews, 2007). Adaptability and flexibility are paramount and, in more practical terms, a lingua franca approach also suggests the need for strategies to deal with variation and diversity. The strategies and skills listed by Seidlhofer (2004: 227) include 'drawing on extralinguistic clues, identifying and building shared knowledge, gauging and adjusting to interlocutors' linguistic repertoires, supportive listening, signalling non-comprehension in a face-saving way, and the like'. To some extent, the use of pair work tasks in assessment provides opportunities to acknowledge these strategies, but much work remains to be done.

In this chapter, I have attempted to show that adopting a lingua franca perspective does not mean rejecting standard polarities of language use, or adopting reified models, whether lingua franca or local. It mainly involves two things. First, it suggests a features-based evaluation of priorities. This is based largely on the criterion of intelligibility (see Lindemann, this volume), combined with a pragmatic awareness of local and global influences. Pronunciation descriptors and examiner guidelines may be in need of modification, in order to acknowledge non-standard features that are unproblematic for intelligibility. Secondly, it suggests the de-emphasizing of system reproduction, in favour of encouraging the skills and strategies that go hand in hand with adaptability. A lingua franca approach acknowledges that the local is global, and vice versa: globalization has made the local a lot more complex, but teaching and assessment still need to take account of local vantage points. Such an approach has concrete indications for pronunciation teaching and assessment in 'Asia's World City', and perhaps elsewhere.

Notes

(1) Phrases such as 'amount of work done' should not of course be taken too literally, as sounds themselves cannot 'do work'; what they can be said to do is to aid processes of meaning construction in the mind of the listener. Statements to the effect that sounds 'do work', while compelling as figures of speech, tend to encourage what Harris (2002) calls the myth of 'telementation' (the idea that communication involves the transmission of messages from one mind to another; see also Croft, 2000, on the 'conduit metaphor').

(2) I do not have detailed information on the examiners, but they include both nonnative and native English speakers and generally have considerable experience of the local context. It might therefore be argued that their impressions of intelligibility were affected by their familiarity with Cantonese and/or Cantonese-accented English (see Browne & Fulcher, this volume). This is certainly possible, but other studies have shown that nonnative listeners tend to be harsher in their judgements

of nonnative accent features (see Hu & Lindemann, 2009). It should also be noted that certain Hong Kong English accent features are capable of reducing intelligibility even for local listeners who are familiar with the local accent (see Sewell, 2015).

References

Abercrombie, D. (1949) Teaching pronunciation. *English Language Teaching* 3, 113–122.

Andrews, S.J. (2007) *Teacher Language Awareness*. Cambridge: Cambridge University Press.

Bardovi-Harlig, K. (2007) One functional approach to second language acquisition: The concept-oriented approach. In B. VanPatten and J. Williams (eds) *Theories in Second Language Acquisition: An Introduction* (pp. 57–75). Mahwah, NJ: Lawrence Erlbaum.

Bassetti, B., Escudero, P. and Hayes-Harb, R. (2015) Second language phonology at the interface between acoustic and orthographic input. *Applied Psycholinguistics* 36 (1), 1–6.

Blommaert, J. and Rampton, B. (2011) Language and superdiversity. *Diversities* 13 (2), 1–21.

Brown, A. (1991) Functional load and the teaching of pronunciation. In A. Brown (ed.) *Teaching English Pronunciation: A Book of Readings* (pp. 221–224). London and New York: Routledge.

Bybee, J. (2001) *Phonology and Language Use*. Cambridge: Cambridge University Press.

Canagarajah, S. (2007) Lingua franca English, multilingual communities, and language acquisition. *Modern Language Journal* 91, 923–939.

Catford, J.C. (1987) Phonetics and the teaching of pronunciation: A systemic description of English phonology. In J. Morley (ed.) *Current Perspectives on Pronunciation: Practices Anchored in Theory* (pp. 87–100). Washington, DC: TESOL.

Chan, J.Y. (2014) An evaluation of the pronunciation target in Hong Kong's ELT curriculum and materials: Influences from WE and ELF? *Journal of English as a Lingua Franca* 3 (1), 145–170.

Coniam, D. (2013) Ten years on: The Hong Kong Language Proficiency Assessment for Teachers of English (LPATE). *Language Testing* 30 (1), 147–155.

Coniam, D. and Falvey, P. (2002) Selecting models and setting standards for teachers of English in Hong Kong. *Journal of Asian Pacific Communication* 12 (1), 13–38.

Cook, V. (2011) Teaching English as a foreign language in Europe. In E. Hinkel (ed.) *Handbook of Research in Second Language Teaching and Learning* (Vol. 2, pp. 140–154). London and New York: Routledge.

Croft, W. (2000) *Explaining Language Change: An Evolutionary Approach*. Harlow: Longman.

Cruttenden, A. (2014) *Gimson's Pronunciation of English*. New York: Routledge.

Cutler, A. (2005) The lexical statistics of word recognition problems caused by L2 phonetic confusion. *Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, 2005* (pp. 413–416).

Cutler, A., Treiman, R. and Van Ooijen, B. (2010) Strategic deployment of orthographic knowledge in phoneme detection. *Language and Speech* 53 (3), 307–320.

Dauer, R. (2005) The lingua franca core: A new model for pronunciation instruction? *TESOL Quarterly* 39 (3), 543–550.

Derwing, T. (2008) Curriculum issues in teaching pronunciation to second language learners. In J.G. Hansen Edwards and L. Zampini (eds) *Phonology and Second Language Acquisition* (pp. 347–369). Amsterdam: John Benjamins.

Deterding, D. (2010) Norms for pronunciation in Southeast Asia. *World Englishes* 29 (3), 364–367.

Deterding, D. (2011) English language teaching and the lingua franca core in East Asia. *Proceedings of the International Conference of Phonetic Sciences* XVII, 92–95.

Deterding, D. (2013) *Misunderstandings in English as a Lingua Franca: An Analysis of ELF Interactions in South-East Asia.* Berlin: Walter de Gruyter.

Dziubalska-Kolaczyk, K. and Przedlacka, J. (eds) (2005) *English Pronunciation Models: A Changing Scene*. Bern: Peter Lang.
Elder, C. and Davies, A. (2006) Assessing English as a lingua franca. *Annual Review of Applied Linguistics* 26, 282–304.
Elder, C. and Harding, L. (2008) Language testing and English as an international language: Constraints and contributions. *Australian Review of Applied Linguistics* 31 (3), 34.1–34.11.
Evans, S. (2011) Hong Kong English and the professional world. *World Englishes* 30 (3), 293–316.
Ferguson, G. (2009) Issues in researching English as a lingua franca: A conceptual inquiry. *International Journal of Applied Linguistics* 19 (2), 117–135.
Field, J. (2003) Promoting perception: Lexical segmentation in L2 listening. *ELT Journal* 57 (4), 325–334.
Field, J. (2004) Pronunciation acquisition and the individual learner. Presentation at the IATEFL Joint Pronunciation and Learner Independence Special Interest Groups Event, University of Reading, 26 June.
Field, J. (2005) Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly* 39 (3), 399–423.
Givón, T. (1995) *Functionalism and Grammar*. Amsterdam: John Benjamins.
Grosjean, F. and Gee, J.P. (1987) Prosodic structure and spoken word recognition. *Cognition* 25, 135–155.
Hall, C.J. (2013) Cognitive contributions to plurilithic views of English and other languages. *Applied Linguistics* 34 (2), 211–231.
Harris, R. (2002) The role of the language myth in the Western cultural tradition. In R. Harris (ed.) *The Language Myth in Western Culture* (pp. 1–24). Richmond: Curzon.
Held, D., McGrew, A., Goldblatt, D. and Perraton, J. (1999) *Global Transformations: Politics, Economics and Culture*. Cambridge: Polity.
HKEAA (Hong Kong Examinations and Assessment Authority) (2003–2009) Language Proficiency Assessment for Teachers (English Language): Assessment reports. See http://www.edb.gov.hk
HKEAA (Hong Kong Examinations and Assessment Authority) (2011) Language Proficiency Assessment for Teachers (English Language): Handbook. See http://www.edb.gov.hk
HKEAA (Hong Kong Examinations and Assessment Authority) (2013) *Hong Kong Diploma of Secondary Education Examination (English Language): Examination Report and Question Papers*. Hong Kong: Hong Kong Government Printer.
Hu, G. and Lindemann, S. (2009) Stereotypes of Cantonese English, apparent native/non-native status, and their effect on non-native English speakers' perception. *Journal of Multilingual and Multicultural Development* 30 (3), 253–269.
Hung, T.T.N. (2000) Towards a phonology of Hong Kong English. *World Englishes* 19 (3), 337–356.
Jenkins, J. (2000) *The Phonology of English as an International Language: New Models, New Norms, New Goals*. Oxford: Oxford University Press.
Katamba, F. (2005) *English Words: Structure, History, Usage* (2nd edn). London and New York: Routledge.
Kohler, K.J. (2011) Communicative functions integrate segments in prosodies and prosodies in segments. *Phonetica* 68 (1–2), 26–56.
Levis, J.M. (2005) Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly* 39 (3), 369–377.
Lin, Y.H. (2003) Interphonology variability: Sociolinguistic factors affecting L2 simplification strategies. *Applied Linguistics* 24 (4), 439–464.
Long, M.H. and Norris, J.M. (2009) Task-based teaching and assessment. In K. Van den Branden, M. Bygate and J. Norris (eds) *Task-based Language Teaching: A Reader* (pp. 135–142). Amsterdam: John Benjamins.

Luce, P.A. and Pisoni, D.B. (1998) Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19 (1), 1–36.
Marslen-Wilson, W. and Zwitserlood, P. (1989) Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance* 15 (3), 576–585.
Mitchell, R., Myles, F. and Marsden, E. (2013) *Second Language Learning Theories*. London and New York: Routledge.
Munro, M.J. (2013) Intelligibility. In C. Chapelle (ed.) *The Encyclopedia of Applied Linguistics* (pp. 1–7). Hoboken, NJ: Wiley-Blackwell.
Munro, M.J. and Derwing, T.M. (2006) The functional load principle in ESL pronunciation instruction: An exploratory study. *System* 34 (4), 520–531.
Munro, M.J. and Derwing, T.M. (2015) A prospectus for pronunciation research in the 21st century: A point of view. *Journal of Second Language Pronunciation* 1 (1), 11–42.
Nunan, D. (2004) *Task-based Language Teaching*. Cambridge: Cambridge University Press.
Ockey, G. and French, R. (2014) From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*. Published online. doi:10.1093/applin/amu060.
Ong, W. (2002) *Orality and Literacy*. London and New York: Routledge.
Parkin, D. (2012) From multilingual classification to translingual ontology: Concluding commentary. *Diversities* 14 (2), 72–85.
Rajadurai, J. (2007) Intelligibility studies: A consideration of empirical and ideological issues. *World Englishes* 26 (1), 87–98.
Schreier, D. (2009) How diagnostic are English universals? In M. Filppula, J. Klemola and H. Paulasto (eds) *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond* (pp. 57–79). London and New York: Routledge.
Seidlhofer, B. (2004) Research perspectives on teaching English as a lingua franca. *Annual Review of Applied Linguistics* 24, 209–239.
Setter, J. (2006) Speech rhythm in world Englishes: The case of Hong Kong. *TESOL Quarterly* 40 (4), 763–782.
Sewell, A. (2013) Language testing and international intelligibility: A Hong Kong case study. *Language Assessment Quarterly* 10 (4), 423–443.
Sewell, A. (2015) The intranational intelligibility of Hong Kong English accents. *System* 49, 86–97.
Sewell, A. and Chan, J.S.C. (2010) Patterns of variation in the consonantal phonology of Hong Kong English. *English World-Wide* 31 (2), 138–161.
Shannon, C. (1951) Prediction and entropy of spoken English. *Bell System Technical Journal* 29, 50–64.
Surendran, D. and Niyogi, P. (2006) Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. In O.N. Thomsen (ed.) *Competing Models of Language Change: Evolution and Beyond* (pp. 43–58). Amsterdam: John Benjamins.
Sussex, R. and Kirkpatrick, A. (2012) A postscript and a prolegomenon. In A. Kirkpatrick and R. Sussex (eds) *English as an International Language in Asia: Implications for Language Education* (pp. 223–231). Dordrecht: Springer.
Swidler, A. (2001) What anchors cultural practices. In T.R. Schatzki, K. Knorr Cetina and E. von Savigny (eds) *The Practice Turn in Contemporary Theory* (pp. 74–92). London and New York: Routledge.
Vertovec, S. (2007) Super-diversity and its implications. *Ethnic and Racial Studies* 30 (6), 1024–1054.
Yazan, B. (2015) Intelligibility. *ELT Journal.* Published online. doi:10.1093/elt/ccu073.
Zhang, L.J. (2003) Extending the reach of middle school EFL teachers in the People's Republic of China. In H.W. Kam and R.Y.L. Wong (eds) *English Language Teaching in East Asia Today* (pp. 147–162). Singapore: Eastern University Press.

Part 5

Concluding Remarks

14 Second Language Pronunciation Assessment: A Look at the Present and the Future

Pavel Trofimovich and Talia Isaacs

Introduction

Over three decades ago, Michael Canale summarized what he considered to be the challenges facing language assessment in the era of communicative language learning and teaching:

> Just as the shift in emphasis from language form to language use has placed new demands on language teaching, so too has it placed new demands on language testing. Evaluation within a communicative approach must address, for example, new content areas such as sociolinguistic appropriateness rules, new testing formats to permit and encourage creative, open-ended language use, new test administration procedures to emphasize interpersonal interaction in authentic situations, and new scoring procedures of a manual and judgemental nature. (Canale, 1984: 79)

Applied to second language (L2) pronunciation assessment, this description remains highly relevant today, raising a number of important issues, such as: broadening the scope of pronunciation assessment beyond the focus on a single aspect of pronunciation (e.g. segmental accuracy) or a single standard (e.g. absence of a discernible nonnative accent); targeting pronunciation assessment for various interlocutors in interactive settings, for instance, outside a typical focus on academic performance by students in Western societies; as well as developing and fine-tuning novel assessment instruments and procedures. Above all, Canale's description aptly summarizes an ongoing quest in language assessment to capture the authenticity and interactiveness of

language use (e.g. Bachman, 1990; Bachman & Palmer, 2010). The contributions to this edited volume address some of the challenges identified by Canale in innovative ways. Before summarizing these contributions, we hasten to add that no edited volume, including this one, can provide an exhaustive overview of all possible issues in L2 pronunciation assessment; most chapters in this volume are focused on testing or informal evaluative judgements of speech in real-world settings and not on classroom-based assessment, including diagnostic assessment or feedback on test takers' performance. However, the range of topics, the variety of research methodologies and paradigms, and the scope of implications featured here make this volume a timely addition to the growing area of L2 pronunciation assessment.

Current Trends

A focus on intelligibility

According to Levis (2005: 370) and echoed in **Harding**'s chapter, teaching and, by extension, assessing L2 pronunciation can be characterized as the tension between two 'contradictory principles'. The nativeness principle holds that nativelike, unaccented pronunciation is both a chief goal of pronunciation learning and a standard for pronunciation assessment. By contrast, the intelligibility principle posits that the primary goal of pronunciation learning is for learners to be understood by their interlocutors, with the consequence that intelligibility, rather than nativeness, emerges as an appropriate assessment criterion. The research findings are clear: a noticeable or even strong nonnative accent does not always involve a lack of understanding (Derwing & Munro, 2015).

While most applied linguists would agree that intelligibility, rather than a native accent, should be considered as the appropriate target of pronunciation teaching and learning, the uptake and implementation of the intelligibility construct in language assessment have seen multiple shortcomings. One example of such limitations comes from **Harding**'s qualitative analyses of focus group discussions targeting raters' experience using the Common European Framework of Reference (CEFR) Phonological control scale to rate L2 pronunciation (Council of Europe, 2001). One of the most telling outcomes of this study is that raters believe the scale to be skewed in its treatment of accented versus understandable speech and also to include erratic descriptions of pronunciation features across scale levels. For instance, while lower levels of the scale make reference to speakers' accent, its higher levels refer to intelligibility as a criterion or exclude reference to either construct altogether. Harding reports that, in operational uses of the scale, raters appear to be 'filling in' gaps in scale descriptors, attempting to balance a focus on accent with the perceived need for speakers to be intelligible. This

is an important finding as it not only highlights possible weaknesses of the CEFR phonological control scale but also illustrates how a scale can be developed and refined through consultations with its end-users (raters). Above all, Harding's research raises important questions about the usability, practicality, and – ultimately – validity of scale-based assessments of L2 pronunciation.

In another chapter, featuring a prominent focus on pronunciation constructs related to listener understanding of L2 speech, **Ballard and Winke** investigate the interplay between speakers' accent and comprehensibility (degree of listeners' understanding) and their acceptability as an ESL teacher, focusing on nonnative listeners. They show that nonnative listeners can easily distinguish between accented speakers and those who sound unaccented. Despite this, nonnative listeners do not seem to readily label accented speakers as unacceptable teachers. Instead, listeners associate speakers' acceptability as a teacher with their perceptions of these speakers' comprehensibility. This finding is important in that it confirms that raters' decisions with real-life consequences might depend more strongly on how easily L2 speech is understood rather than on how unaccented it sounds, echoing previous work by Derwing and Munro (2009), which showed a similar result for nonnative English speaking engineers in an English-medium workplace setting.

A focus on language

If listener understanding, whether termed intelligibility or comprehensibility, is an important assessment criterion, then identifying linguistic barriers to communication can help researchers and teachers isolate pronunciation elements to target in teaching and assessment. A vibrant area of research is the relationship between L2 speakers' comprehensibility, frequently operationalized as the extent of listeners' perceived ease or difficulty of understanding L2 speech as measured using a Likert-type rating scale, as in the **Ballard and Winke** study, and linguistic features that characterize their speech, with the goal of helping teachers, learners and language testers to isolate and then focus on features that are most important for listeners' understanding.

Illustrating this line of research, the chapter by **Saito, Trofimovich, Isaacs and Webb** examines a range of linguistic dimensions which could contribute to listeners' judgements of comprehensibility and which, by extension, could elucidate the properties of the speech that listeners (raters) tend to take into account in their scoring, hence enhancing our understanding of the L2 comprehensibility construct. This study is innovative in that it broadens the scope of linguistic factors linked to comprehensibility to include various lexical dimensions of L2 speech, including lexical polysemy, diversity and appropriateness, as well as morphological accuracy. Comprehensibility emerges as a complex construct, spanning various dimensions of speech,

with the consequence that the teaching and assessment of comprehensible L2 speech should consider not only pronunciation and fluency aspects of speech but also its lexical content, such as the use of appropriate and diverse vocabulary. The extent to which lexical features are sensitive to differences in L2 learners' comprehensibility scores across task type (Crowther *et al.*, 2015) also requires further exploration.

In another study focusing on language, **Galaczi, Post, Li, Barker and Schmidt** target rhythm, one dimension of speech prosody, investigating the extent to which several measures of rhythm could distinguish L2 pronunciation levels for learners from different language backgrounds across the CEFR language proficiency scale (Council of Europe, 2001). This study is a welcome contribution to research on L2 pronunciation learning and assessment because it shows that micro-level measures of rhythm, such as speech rate and duration differences between stressed and unstressed syllables, while being useful overall, might not be precise enough to distinguish fine-grained prosodic differences between adjacent levels of the CEFR scale. This finding adds to a growing body of research in language assessment (Isaacs *et al.*, 2015; Iwashita *et al.*, 2008; Kang, 2013; Kang & Wang, 2014) suggesting that various linguistic measures of L2 pronunciation often fail to distinguish between adjacent levels in multi-level pronunciation scales. And because such scales often rely on listener judgements, this finding raises a related question of how well listeners distinguish fine-grained linguistic differences, especially when using scales featuring seven or more levels.

A focus on pronunciation standards

One of the core issues in L2 assessment concerns the standards or criteria by which various aspects of L2 speech are assessed. As discussed previously, intelligible L2 pronunciation – as distinct from L2 pronunciation that sounds unaccented – is typically considered to be an appropriate reference for both teaching and assessment because it reflects what is important for communication, that is, speakers' ability to be understood by interlocutors (Derwing & Munro, 2015). Nevertheless, for many language learners and teachers, what sounds like native and accent-free pronunciation remains an important teaching and learning goal (Scales *et al.*, 2006; Subtirelu, 2013; Young & Walsh, 2010).

Several chapters in this volume focus on the issue of appropriate standards and norms for L2 pronunciation assessment. In a delightful chapter, which reads as an armchair conversation with the author, **Davies** problematizes the concept of the native speaker, with reference to the assessment of L2 pronunciation, touching upon such topics as a standard language, accent prestige, and discrimination based on accent. An insightful chapter by **Lindemann** takes these ideas further, discussing the highly variable and therefore elusive nature of 'standard' pronunciation by native speakers.

Lindemann convincingly argues that classifying nonnative speech as being 'standard' or not is highly problematic, at least in part because of listeners' expectations about L2 speech and their often biased perceptions of it (e.g. Kang & Rubin, 2009). She concludes that a deficit-based approach to the teaching and assessment of L2 pronunciation – one based on defining specific speech patterns in terms of 'errors' or deviations from what is expected in a standard norm – is indefensible, calling for language testers to incorporate the construct of the listener into assessment instruments while also trying to minimize any potential listener-based biases.

In two related chapters, both Sewell and Kennedy *et al.* discuss lingua-franca intelligibility as a criterion for L2 pronunciation assessment in situations when one or more interlocutors from different linguistic and cultural backgrounds share a common language. **Sewell** conceptualizes lingua franca intelligibility within a broad functionalist view of language, implying that the linguistic elements most relevant to intelligibility are those that tend to carry the most information in communication (e.g. consonant contrasts tend to do more 'work' in communication, compared to vowel contrasts). He illustrates this approach to intelligibility using the case of English in Hong Kong, arguing for a teaching and assessment criterion that is rooted in intelligibility but informed by local, contextual considerations specific to sociocultural realities of language use. To cite Sewell, '[t]he lingua franca approach acknowledges that the local is global, and vice versa'. In a conceptually related chapter, **Kennedy, Blanchet and Guénette** rely on verbal reports to understand teacher-raters' judgements of L2 speech in the context of using French as a lingua franca in Quebec, Canada. They conclude that teacher-raters show considerable variability in the extent to which they place importance on mutual understanding in lingua franca interactions while evaluating their students' pronunciation. Kennedy *et al.* speculate that individual differences across teachers in their formal training in phonetics and phonology, their teaching experience and their own language learning histories might explain their preference for native speaker versus lingua franca models in evaluating their learners' L2 pronunciations. These researchers conclude with a call for more research into teachers' beliefs about language and communication, so that classroom assessments and pedagogical decisions can be understood in the context of teacher cognitions (e.g. Baker, 2014).

A focus on other L2 skills

Three contributions to the current volume illustrate that the assessment of L2 pronunciation has much to learn from the expertise in assessment of other language skills and components. In a chapter focusing on speech fluency, **Browne and Fulcher** eloquently argue for the importance of considering listeners' and raters' familiarity with L2 speech in assessment of L2 pronunciation, including intelligibility (operationalized through a gap-fill

task) and speech delivery (fluency). Through the use of Rasch analyses, which allow for a simultaneous treatment of both raters' and speakers' performances on the same arithmetic (logit) scale, they show that a measure of L2 intelligibility and a scored measure of speech delivery based on a five-point TOEFL iBT scale (Educational Testing Service, 2009) predictably vary as a function of rater familiarity with L2 speech. These findings reinforce the idea that various constructs subsumed by the umbrella term 'L2 pronunciation', including speech delivery (fluency) and intelligibility, are not simply tied to speakers' performance but also reflect specific characteristics of individual listeners. The study also brings to light the issue that ideally in L2 pronunciation research, listener familiarity effects, when not directly the source of investigation, should be controlled for, although this is difficult to implement in practice. One implication for high-stakes testing settings could be that accredited examiners should be screened for factors such as their degree of familiarity with the accented speech of the test takers (Winke *et al.*, 2013), although it is unclear how this could be put into practice in contexts where test takers from numerous language backgrounds are being assessed.

Working in the field of L2 writing, **Knoch** provides a comprehensive 'roadmap' for various issues in assessing L2 writing, including the development and validation of rating scales, effects of raters and tasks on assessment outcomes, and applications for classroom-based assessment. Knoch's summary is valuable; it not only offers a wealth of evidence-based information from a skill that has benefited from a larger volume of language assessment research, pioneering many of the advancements in, for example, rater training (e.g. Weigle, 1998), but it also highlights current gaps in the assessment of L2 pronunciation. This includes a paucity of research on the development and validation of L2 pronunciation rating scales with an adequately operationalized construct, the need for more research-based evidence for task and listener effects, and the dearth of research into classroom-based pronunciation assessment and self-assessment, as well as interactive and paired assessments of L2 pronunciation.

In a chapter focusing on assessment of L2 listening, **Wagner and Toth** critically examine the extent to which authentic and simplified (scripted) listening comprehension materials are appropriate as assessment materials. A survey of L2 test takers who took either authentic or scripted listening comprehension materials clearly shows that L2 users are aware of important differences across these recorded stimuli, for example, rating scripted materials lower in authenticity and naturalness and being aware that scripted materials include clearer pronunciation and fewer hesitation markers. Wagner and Toth persuasively argue for the use of testing materials that illustrate authentic, natural and representative uses of real-world spoken language if the goal of teaching, learning and assessment is for learners to comprehend authentic L2 speech. This research reminds L2 teachers, researchers and test developers

to consider the issues of authenticity when designing and validating L2 listening and pronunciation tasks.

A focus on individual differences

Research on L2 development of various language skills clearly shows that there are often pronounced differences across individual learners in rates and outcomes of L2 learning (DeKeyser, 2012). L2 pronunciation learning is no exception. For instance, the learning of various linguistic dimensions of L2 speech has been linked to learners' age (Abrahamsson & Hyltenstam, 2009), the quantity and quality of their contact with the L2 (Moyer, 2011), their motivation and cultural sensitivity (Alvord & Christiansen, 2012; Baker-Smemoe *et al.*, 2014; Hardison, 2014), and their willingness to communicate (Baran-Łucarz, 2014; Derwing *et al.*, 2008). These findings clearly argue against a 'one-size-fits-all' approach to pronunciation teaching and assessment, suggesting that different learners can respond differently to the same testing materials and procedures, and that different materials and procedures might be necessary for assessment of diverse populations of learners. In a novel study, **Mora and Darcy** focus on these issues by investigating the relationship between three cognitive variables (attention control, phonological short-term memory, and inhibitory control) and L2 learners' performance on several acoustic and rated measures of their L2 pronunciation. More importantly, the participants in this study are two groups of English language learners – monolingual speakers (monolingual Spanish speakers) and functionally bilingual language users (Spanish-Catalan bilinguals). Mora and Darcy report a complex set of findings, showing links between learners' attention control and phonological memory and their English vowel production, but only for the group of monolingual Spanish learners of English. The researchers speculate that individual differences in L2 users' cognitive capacities can influence how specific learner groups perform in particular assessment tasks and with particular types of assessment materials, calling for more investigations into individual learner differences to better understand contributors to the variability of learners' L2 pronunciation performance.

Future Directions

To go back to Canale's quote from 30 years ago, it is fair to say that language researchers and assessment specialists have made some as yet limited empirical inroads into the assessment of L2 pronunciation, enhancing our understanding of the constructs under investigation and developing and validating novel assessment procedures. A case in point is the recent launch of fully automated speaking tests into the competitive market of standardized language testing products, including Person's Versant tests, Pearson's speaking

component of the PTE Academic (Bernstein *et al.*, 2010) and Educational Testing Service's TOEFL iBT patented automatic speech recognition technology used with the TOEFL iBT practice speaking test, SpeechRater (Zechner *et al.*, 2009). These instruments, the first two of which tend to be used for high-stakes purposes (e.g. a language proficiency certification test for pilots), are scored using automated speech recognition algorithms optimized to predict human scoring using acoustic and temporal correlates of auditory pronunciation measures in addition to machine scored measures of other linguistic phenomena. Concerns within the assessment community have been raised about automated assessments of speech due to the machine scoring system's ability, as yet, only to cope with highly predictable L2 speaking tasks (e.g. Chun, 2008), as opposed to discourse-level extemporaneous speaking tasks that elicit more varied interactional patterns (see Isaacs, 2016). Technology is rapidly improving. However, speech recognition programmers need to be steered away from targeting accent reduction by modelling acoustic phenomena that are easy for the machine to score and, instead, prioritize the linguistic factors that are most consequential for intelligibility.

Despite these and similar technological advances and developments in conceptual thinking, there is ample room for future research to enhance our understanding of the processes and outcomes of pronunciation testing. At a practical level, research into the assessment of pronunciation in languages other than English is virtually non-existent (for a rare exception, see Kennedy *et al.*, this volume), and assessment research targeting multilingual lingua franca L2 users in non-Western, non-academic contexts is lacking. Also limited is research targeting the assessment of sociolinguistic and pragmatic functions of L2 pronunciation, and research incorporating nonnative pronunciation models and standards in assessments. With respect to practical implications of assessment research, in a climate where assessment for learning, formative assessment, learning-oriented assessment and dynamic assessment (in contrast to large-scale testing) is gaining currency in promoting teaching and learning (Turner & Purpura, 2016), it would similarly be important to expand research on classroom-based assessment, including the instructional effectiveness of incidental form-focused instruction (i.e. corrective feedback) on L2 pronunciation development (e.g. Lee & Lyster, 2016; Saito & Lyster, 2012). In addition, the ground is fertile to build on preliminary work regarding learners' self-assessment of pronunciation (Dlaska & Krekeler, 2008; Lappin-Fortin & Rye, 2014), including helping learners calibrate their perceptions to those of their interlocutors, thus minimizing distorted perceptions of their speech (Trofimovich *et al.*, 2016).

At the conceptual level, Canale's (1984: 79) call for new testing instruments and procedures involving 'interpersonal interaction in authentic situations' has largely not been answered, emphasizing the need for more research into interactional paired and group assessments involving an L2 pronunciation component. There has been some preliminary research in this area in recent

years using the Cambridge interactional (collaborative) test tasks in research settings (Isaacs, 2013; Jaiyote, 2015), although future research needs to go further in investigating pronunciation features that account for communication breakdowns specifically and discrepancies in the extent to which interlocutors report understanding one another. Last but not least, more theorizing is needed targeting possible models or theories that can serve as conceptual bases for the assessment of L2 pronunciation. For instance, as Isaacs (2014) argues, models of communicative competence (Bachman, 1990; Bachman & Palmer, 1996) are insufficiently nuanced to capture all of the complexities of pronunciation, particularly in relation to pronunciation perception, production, and (where applicable) orthographic effects (see also Fulcher, 2003). There is a further need to clarify the role of holistic pronunciation-related constructs such as intelligibility in relation to more discrete L2 speech measures and, if possible, to listener/rater/interlocutor variables. Therefore, more theory building is required to understand the nature of the phenomena being targeted through assessment and, specifically, to better understand major global constructs in L2 pronunciation so they can be better operationalized in assessment instruments (Isaacs & Trofimovich, 2012; see Foote & Trofimovich, submitted, for a discussion of various theoretical frameworks of L2 pronunciation learning).

We conclude this chapter (and in fact the entire volume) with a list of possible issues and questions that we consider to be important for future research into L2 pronunciation assessment. Clearly, this list is non-exhaustive, yet in our opinion it identifies several priority research axes which, if followed, have the greatest potential for enhancing both the breadth and depth of our understanding of L2 pronunciation assessment.

- How do different stakeholders perceive assessments of pronunciation in formal and informal contexts? In what ways can technology be used to validate listener perceptions of linguistic phenomena?
- What is the effect of individual differences in listeners' cognitive or attitudinal variables on listeners' (raters') judgements of L2 pronunciation and on speakers' communicative success in real-world settings?
- How can sources of construct-irrelevant variance related to listener background variables (e.g. accent familiarity effects) be mitigated in high-stakes assessments of L2 speech? What are the implications for rater screening and training and for mitigating sources of bias?
- Which pronunciation features should be prioritized in L2 pronunciation instruction and assessment? How can these features feed into the development of valid speaking assessment instruments?
- How can measures of pronunciation and fluency normally used for individual learners in lab-based research settings be adapted for use in naturalistic settings, including in conversational interactions with interlocutors? Similarly, how can stimuli used in lab-based settings be adapted to generate more authentic testing prompts (e.g. Jones, 2015)?

- In light of the current debate on the native speaker standard and the coexistence of multiple varieties of English, what is the appropriate standard or language varieties that learners should be exposed to for listening tests, including audio prompts for integrated test tasks (e.g. Ockey & French, 2014)? For example, could Cook's (1992, 2012) construct of multicompetence form the basis of a target language assessment standard that draws on descriptions of proficient multicompetent learners or test takers rather than native speakers (e.g. Brown, 2013)?
- If intelligibility is valued as an assessment criterion by the applied linguistics community, then how can intelligibility feed into models of communicative competence (Canale & Swain, 1980) or communicative language ability (Bachman, 1990)? Should intelligibility be instructed and assessed in conjunction with pragmatic competence, focusing on utterances that are not only clearly understood, but are also pragmatically appropriate in the context of language use (e.g. Yates, 2014)?
- How can findings on form-focused instruction in L2 learning and teaching, on the instructional effectiveness of pronunciation interventions, including corrective feedback, and longitudinal studies on learner pronunciation development over time, inform formative assessment practices, particularly in classroom settings?
- How can we move beyond Lado (1961), taking into account technological advancements and the latest developments in research and pedagogy, to bring pronunciation assessment out of its time warp and integrate it into mainstream assessment research and practice?
- To parallel calls to foster language educators' assessment literacy (e.g. Fulcher, 2012; Taylor, 2009), how can we improve experienced teachers' and assessment researchers' and practitioners' pronunciation literacy, making it more mainstream and accessible?

References

Abrahamsson, N. and Hyltenstam, K. (2009) Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning* 59, 249–306.

Alvord, S.M. and Christiansen, D.E. (2012) Factors influencing the acquisition of Spanish voiced stop spirantization during an extended stay abroad. *Studies in Hispanic and Lusophone Linguistics* 5, 239–276.

Bachman, L.F. (1990) *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L.F. and Palmer, A.S. (1996) *Language Testing in Practice: Designing and Developing Useful Language Tests.* Oxford: Oxford University Press.

Bachman, L. and Palmer, A (2010) *Language Assessment in Practice.* Oxford: Oxford University Press.

Baker, A. (2014) Exploring teachers' knowledge of second language pronunciation techniques: Teacher cognitions, observed classroom practices, and student perceptions. *TESOL Quarterly* 48, 136–163.

Baker-Smemoe, W., Dewey, D.P., Bown, J. and Martinsen, R.A. (2014) Variables affecting L2 gains during study abroad. *Foreign Language Annals* 47, 464–486.
Baran-Łucarz, M. (2014) The link between pronunciation anxiety and willingness to communicate in the foreign-language classroom: The Polish EFL context. *Canadian Modern Language Review* 70, 445–473.
Bernstein, J., Van Moere, A. and Cheng, J. (2010) Validating automated speaking tests. *Language Testing* 27, 355–377.
Brown, A. (2013) Multicompetence and second language assessment. *Language Assessment Quarterly*, 10, 219–235.
Canale, M. (1984) Testing in a communicative approach. In G.A. Jarvis (ed.) *The Challenge for Excellence in Foreign Language Education* (pp. 79–92). Middlebury, VT: Northeast Conference for the Teaching of Foreign Languages.
Canale, M. and Swain, M. (1980) Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1, 1–57.
Chun, C.W. (2008) Comments on 'evaluation of the usefulness of the *Versant for English* test: A response': The author responds. *Language Assessment Quarterly* 5 (2), 168–172.
Cook, V.J. (1992) Evidence for multicompetence. *Language Learning* 42 (4), 557–591.
Cook, V. (2012) Multi-competence. In C.A. Chapelle (ed.) *The Encyclopedia of Applied Linguistics* (pp. 3768–3774). Oxford: Wiley-Blackwell.
Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
Crowther, D., Trofimovich, P., Isaacs, T. and Saito, K. (2015) Does a speaking task affect second language comprehensibility? *Modern Language Journal* 99, 80–95.
DeKeyser, R. (2012) Interactions between individual differences, treatments, and structures in SLA. *Language Learning* 62, 189–200.
Derwing, T.M. and Munro, M.J. (2009) Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. *Canadian Modern Language Review* 66, 181–202.
Derwing, T.M. and Munro, M.J. (2015) *Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research*. Amsterdam: John Benjamins.
Derwing, T.M., Munro, M.J. and Thomson, R.I. (2008) A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics* 29, 359–380.
Dlaska, A. and Krekeler, C. (2008) Self-assessment of pronunciation. *System* 36, 506–516.
Educational Testing Service (2009) *The Official Guide to the TOEFL Test* (3rd edn). New York: McGraw-Hill.
Foote, J.A. and Trofimovich, P. (in press) Second language pronunciation learning: An overview of theoretical perspectives. In O. Kang, R.I. Thomson and J. Murphy (eds) *The Routledge Handbook of Contemporary English Pronunciation*. London: Routledge.
Fulcher, G. (2003) *Testing Second Language Speaking*. London: Pearson.
Fulcher, G. (2012) Assessment literacy for the language classroom. *Language Assessment Quarterly* 9 (2), 113–132.
Hardison, D.M. (2014) Changes in second-language learners' oral skills and socio-affective profiles following study abroad: A mixed-methods approach. *Canadian Modern Language Review* 40, 415–444.
Isaacs, T. (2013) International engineering graduate students' interactional patterns on a paired speaking test: Interlocutors' perspectives. In K. McDonough and A. Mackey (eds) *Second Language Interaction in Diverse Educational Settings* (pp. 227–246). Amsterdam: John Benjamins.
Isaacs, T. (2014) Assessing pronunciation. In A.J. Kunnan (ed.) *The Companion to Language Assessment* (pp. 140–155). Hoboken, NJ: Wiley-Blackwell.

Isaacs, T. (2016) Assessing speaking. In D. Tsagari and J. Banerjee (eds) *Handbook of Second Language Assessment* (pp. 131–146). Berlin: DeGruyter Mouton.

Isaacs, T. and Trofimovich, P. (2012) 'Deconstructing' comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition* 34, 475–505.

Isaacs, T., Trofimovich, P., Yu, G. and Chereau, B.M. (2015) Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS Research Reports Online* 4, 1–48.

Iwashita, N., Brown, A., Mcnamara, T. and O'Hagan, S. (2008) Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics* 29 (1), 24–49.

Jaiyote, S. (2015) The relationship between test-takers' L1, their listening proficiency and their performance in pairs. A*RAGs Research Reports Online, AR-A/2015/2*. Manchester: British Council.

Jones, J. (2015) Exploring open consonantal environments for testing vowel perception. Unpublished Master's thesis, University of Melbourne.

Kang, O. (2013) Linguistic analysis of speaking features distinguishing general English exams at CEFR levels. *Research Notes* 52, 40–48.

Kang, O. and Rubin, D.L. (2009) Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology* 28, 441–456.

Kang, O. and Wang, L. (2014) Impact of different task types on candidates' speaking performances and interactive features that distinguish between CEFR levels. *Research Notes* 57, 40–49.

Lado, R. (1961) *Language Testing: The Construction and Use of Foreign Language Tests.* London: Longman.

Lappin-Fortin, K. and Rye, B.J. (2014) The use of pre-/posttest and self-assessment tools in a French pronunciation course. *Foreign Language Annals* 47 (2), 300–320.

Lee, A.H. and Lyster, R. (2016) Effects of different types of corrective feedback on receptive skills in a second language: A speech perception training study. *Language Learning*. Published online. doi:10.1111/lang.12167.

Levis, J.M. (2005) Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly* 39, 369–377.

Moyer, A. (2011) An investigation of experience in L2 phonology. *Canadian Modern Language Review* 67, 191–216.

Ockey, G. and French, R. (2014) From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*. Published online. doi:10.1093/applin/amu060.

Saito, K. and Lyster, R. (2012) Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/ by Japanese learners of English. *Language Learning* 62, 595–633.

Scales, J., Wennerstrom, A., Richard, D. and Wu, S.H. (2006) Language learners' perceptions of accent. *TESOL Quarterly* 40, 715–738.

Subtirelu, N. (2013) What (do) learners want (?): A re-examination of the issue of learner preferences regarding the use of 'native' speaker norms in English language teaching. *Language Awareness* 22, 270–291.

Taylor, L. (2009) Developing assessment literacy. *Annual Review of Applied Linguistics*, 29, 21–36.

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K. and Crowther, D. (2016) Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition* 19, 122–140.

Turner, C.E. and Purpura, J.E. (2016) Learning-oriented assessment in the classroom. In D. Tsagari and J. Banerjee (eds) *Handbook of Second Language Assessment* (pp. 255–274). Berlin: DeGruyter Mouton.

Weigle, S.C. (1998) Using FACETS to model rater training effects. *Language Testing* 15, 263–287.

Winke, P., Gass, S. and Myford, C. (2013) Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing* 30, 231–252.

Yates, L. (2014) Learning how to speak: Pronunciation, pragmatics and practicalities in the classroom and beyond. *Language Teaching*. Published online. doi.org/10.1017/S0261444814000238.

Young, T.J. and Walsh, S. (2010) Which English? Whose English? An investigation of 'non-native' teachers' beliefs about target varieties. *Language, Culture, and Curriculum* 23, 123–137.

Zechner, K., Higgins, D., Xi, X. and Williamson, D.M. (2009) Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication* 51, 883–895.

Index